Modeling synchronization effects in the yeast cell cycle

Dissertation
Zur Erlangung des akademischen Grades
Doctor rerum naturalium
(Dr. rer. nat.)

im Fach
Biophysik

eingereicht an der
Lebenswissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

von
M. Sc. Julia Katharina Schlichting

Präsidentin der Humboldt-Universität zu Berlin
Prof. Dr.-Ing. habil. Dr. Sabine Kunst
Dekan der Lebenswissenschaftlichen Fakultät
Prof. Dr. Bernhard Grimm

Gutachter/innen:
1. Prof. Dr. Dr. h.c. Edda Klipp
2. Prof. Dr. Hanspeter Herzel
3. Prof. Dr. Jens Timmer

Tag der mündlichen Prüfung: 18.03.2019

Why do we study yeast cells?

At the beginning of my PhD, I was complaining that studying yeast cells will never be beneficial for human beings. It took me three years to understand that especially cancer research benefits from studying the yeast cell cycle. Even if my study may not save people, I found a deeper sense of my work.

Zusammenfassung

Einer der bekanntesten Modellorganismen in der Systembiologie ist die Bäckerhefe, deren wissenschaftlicher Name *Saccharomyces cerevisiae* ist. Er wird oft verwendet, um den mitotischen Zellzyklus eukaryotischer Zellen zu erforschen. Der Zellzyklus ist ein komplexer Prozess, dessen Regulation über Cycline, Cyclin-abhängige Kinasen (CDK) und CDK-Inhibitoren (CKI) erfolgt. Cdc28 ist die wichtigste Kinase des Zellzyklus. An verschiedenen Kontrollpunkten innerhalb des Zellzyklus entscheidet die Zelle, ob der Zellzyklus fortgesetzt wird oder nicht. Der wichtigste Kontrollpunkt reguliert den Übergang von der G1 in die S Phase und entscheidet daher, ob die Zelle eine weitere Zellteilung durchläuft. Desshalb nennt man diesen Kontrollpunkt auch START.

Im Rahmen dieser Arbeit verwenden wir einen stochastischen Modellierungsansatz, um die Auswirkungen verschiedener Synchronisationsmethoden auf den Zellzyklus zu untersuchen. Des weiteren interessiert uns, welche Unterschiede zwischen unsynchronisierten und synchronisierten Zellen bestehen. Um entsprechende Modellparameter zu schätzen, kombinieren wir Phasen aufgelöste mRNA-Verteilungen und Protein-Zeitreihen. Die Phasen aufgelösten mRNA-Verteilungen wurden für unsynchronisierte Einzelzellen bestimmt. Hingegen basieren die Protein-Zeitreihen auf synchronisierten Zellpopulationen. Auf diese Weise können wir mRNA-Dynamiken für ausgewählte Synchronisationsmethoden vorhersagen. Wir benutzen einen zweistufigen Optimierungsansatz, in dem wir zwischen mRNA- und Protein-Ebene unterscheiden. Die Parameterschätzung selbst basiert auf der Maximum-Likelihood-Methode. Unter der Verwendung der smFISH-Technik[1] haben wir Phasen aufgelösten mRNA-Verteilungen für drei mRNA-Spezies gemessen: *SIC1*, *CLN2* und *CLB5*. Die Protein-Zeitreihen wurden mit Hilfe von Western Blots für drei Protein-Spezies gemessen: Sic1, Cln2 und Clb5. Bei den gemessenen Molekülen handelt es sich um die Hauptregulatoren des G1-S Phasenübergangs, welche die Komponenten unseres Zellzyklusmodells darstellen.

Durch die erfolgreiche Integration von qualitativ unterschiedlichen Datentypen in der Parameterschätzung konnten wir erstmals eine systematische Analyse von Synchronisationseffekten auf den Zellzyklus durchführen. Der Unterschied von synchronisierten zu unsynchronisierten Zellen besteht hauptsächlich darin, dass der zeitliche Ablauf des Zellzyklus verändert ist. Die stärksten zeitlichen Veränderungen weist die Synchronisation mit α-Faktor auf. Elutrierte Zellen sind den unsynchronisierten Zellen am ähnlichsten, auch wenn diese eine deutlich verlängerte G1 Phase aufweisen. Wir zeigen in dieser Arbeit, dass synchronisierte Zellpopulationen unzureichend sind, um Rückschlüsse auf den Zellzyklus unsynchronisierter Zellen zu ziehen.

[1] smFISH ist eine Abkürzung aus dem Englischen und steht für „single molecule RNA *in situ* hybridization“.

Abstract

Saccharomyces cerevisiae is a famous model organism in systems biology to study the mitotic cell cycle in eukaryotic cells. The cell cycle is a highly controlled process which is regulated by cyclins, cycline-dependent kinases (CDK) and cyclin-dependent kinase inhibitors (CKI). The main kinase involved in cell cycle regulation is Cdc28. START is the most important check point and controls the G1 to S phase transition. At this point, cells decide if they enter a new cell division cycle or not.

In this study, we analyze influences of different synchronization methods on the cell cycle and differences between unsynchronized and synchronized cells by using a stochastic modeling approach. We combine phase-resolved mRNA distributions of unsynchronized single cells and protein time courses of synchronized cell populations to estimate model parameters and to predict synchronization specific mRNA dynamics. Parameter estimation is based on a maximum likelihood approach and performed in a 2-step-optimization in which we differentiate between mRNA and protein level. We measured phase-resolved mRNA distributions of mRNA species *SIC1*, *CLN2* and *CLB5* by smFISH[2] and protein time courses of protein species Sic1, Cln2 and Clb5 by Western blotting. These molecules are key regulators of the G1 to S phase transition and represent components of our cell cycle model.

By integrating qualitatively different data types in parameter estimation, we come up with a systematic analysis of synchronization effects on the cell cycle. Cell cycle timing is mainly responsible for differences between unsynchronized and synchronized cells and is mostly affected in α-factor synchronized cells. Ignoring the prolongation of the G1 phase, elutriated cells are most similar to unsynchronized cells. We show that synchronized cell populations are insufficient to derive general cell cycle behavior of unsynchronized cells.

[2]smFISH is an abbreviation for single molecule RNA *in situ* hybridization.

Contents

I Introduction

1 Scientific framework . . . 15
1.1 Yeast as a model organism . . . 15
1.2 Studying the mitotic cell cycle . . . 15
1.3 Synchronizing cell populations . . . 16
1.4 Measuring single cell and population data . . . 17
1.5 Mathematical modeling in systems biology . . . 17

2 General interest of this study . . . 19

II Methods

3 Experimental data . . . 23
3.1 Counting mRNA molecules in single cells by smFISH . . . 23
3.2 Measuring relative protein abundances of cell populations by Western blotting . . . 23
3.3 Pre-processing Western blot data . . . 24
3.4 Normalizing Western blot data by using absolute protein numbers from the PaxDb database . . . 24

4 Mathematical modeling . . . 27
4.1 Modeling the chemical reaction system . . . 27
4.2 Formulating the chemical master equation . . . 28
4.3 Transition to the reaction rate equation . . . 31
4.4 Using Gillespie's stochastic simulation algorithm . . . 35
4.5 Representation of the experimental data by the stochastic model . . . 41

5 Parameter estimation . . . 45
5.1 Estimating parameters in ODE systems . . . 45
5.2 Applying the maximum likelihood approach . . . 46
5.3 Combining count and time series data in a 2-step-optimization . . . 49
5.4 Performing global optimization by multi-start local optimization . . . 49

6 Identifiability analysis . . . 51
6.1 Introducing the concept of parameter identifiability . . . 51
6.2 Working with profile likelihoods . . . 51
6.3 Classification of protein profile effects . . . 54

7 L_1 regularization . . . 57
7.1 Regularizing the likelihood function . . . 57
7.2 Identifying protein parameters not covered by the data . . . 58
7.3 Calculating fold changes between mRNA parameters of unsynchronized and synchronized cells . . . 59

III Results

8 Experimental data . . . 63
8.1 smFISH data reveal low numbers of mRNA molecules . . . 63
8.2 Western blot data show differences in protein abundances between synchronization methods . . . 63
8.3 Pre-processing indicates synchronization specific protein numbers . . . 66

9 Mathematical modeling . . . 67
9.1 Synchronization influences cell cycle timing and gene transcription . . . 67
9.2 Observables successfully reproduce protein time courses of different synchronization methods . . . 69

10 Parameter estimation . . . 71
10.1 mRNA optimization perform superior to protein optimization . . . 71
10.2 Parameters related to cell cycle timing are most clearly affected by synchronization 71

11 Identifiability analysis . . . 77
11.1 mRNA parameters are generally identifiable . . . 77
11.2 Cell cycle timing and gene transcription are equally affected by Western blot data 77
11.3 Protein parameters are predominantly practically non-identifiable . . . 79
11.4 mRNA priors contribute less to protein parameters of α-factor synchronized cells . . 80
11.5 Parameter dependencies persist between identifiable and non-identifiable parameters . . . 82
11.6 Variations in model trajectories occur equally for identifiable and non-identifiable parameters . . . 83

12 L_1 regularization . . . 89
12.1 L_1 regularization improves identifiability of protein parameters . . . 89
12.2 mRNA fold changes suggest synchronization specific smFISH measurements . . . 89

IV Discussion

13 Discussion . . . 95
13.1 Discussing method specific results . . . 95
13.2 Regarding general interests of this study . . . 98

14 Final statement . . . 101

V Appendix

A Symbol directory . . . 105

B Abbreviations . . . 111

C Definitions . . . 113
C.1 Definition of time points t_{ln} . . . 113
C.2 Definition of the number of cells $\mathcal{Z}_{kl}$ . . . 113
C.3 Definition of the number of mRNA molecules A_{ka} . . . 113

D Calculations **115**
D.1 Calculation of pre-processed data by using blotIt2 115
D.2 Calculation of the time evolution of the expected state 115
D.3 Calculation of the time until next reaction for time dependent stochastic rates . . . 116
D.4 Calculation of the stationary distribution of a birth-death process 116
D.5 Calculation of the analytical solution of the CME for a birth-death process with a Poisson initial distribution 117
D.6 Calculation of the gradient of the objective function in the mRNA optimization step 118
D.7 Calculation of the gradient of the objective function in the protein optimization step 119
D.8 Calculation of time dependent mRNA production rates by using Fermi-Dirac distributions 119
D.9 Calculation of p-values in a χ^2 goodness of fit test 120
D.10 Latin hypercube sampling versus Gaussian sampling 120

E Figures **121**

F Tables **155**

Bibliography **163**

Acknowledgement **171**

Declaration of authorship **173**

Introduction

1 Scientific framework 15

1.1 Yeast as a model organism
1.2 Studying the mitotic cell cycle
1.3 Synchronizing cell populations
1.4 Measuring single cell and population data
1.5 Mathematical modeling in systems biology

2 General interest of this study 19

1. Scientific framework

1.1 Yeast as a model organism

Saccharomyces cerevisiae, also known as Baker's yeast or budding yeast, is among the best studied experimental organism [1]. It is the same yeast people have been using for thousands of years to brew beer or bake bread [2]. In 1996, its genome was sequenced as the first eukaryotic organism [3]. The genome[3] comprises 12157105 base pairs and encodes 7036 genes with 6600 coding genes, 424 non-coding genes and 12 pseudogenes [4]. Yeast is a single cell organism, is very small (30-50 μm^3) [5], has a limited number of crucial molecules (5-10 mRNA copies, 500-5000 protein copies) [5], has a short generation time of about 90 minutes [6] and is easy to cultivate. Further, genetic manipulations are cheap compared to more complex systems.

Eukaryotic cells are characterized by membrane-bound organelles. The most important organelle is the nucleus in which DNA is packed in chromosomes. Yeast has 16 linear chromosomes [6]. A number of biological processes are shared between human and yeast cells, as is the mitotic cell cycle. About 20% of human genes involved in diseases, e.g. cancer, have counterparts in yeast cells[4]. Cancer come along with disfunctioning of the cell cycle why cancer research essentially benefits from studying the yeast cell cycle. In 2002, Leland H. Hartwell, Tim Hunt and Sir Paul M. Nurse received the Nobel Prize in physiology or medicine due to their discoveries of key regulators of the yeast cell cycle [7].

1.2 Studying the mitotic cell cycle

The mitotoc cell cycle describes the change of DNA replication, chromosome segregation and cell division of diploid cells (2N) [8, 9]. It is composed of an interphase and mitosis (see Figure 1.1). The interphase comprises two gap phases (G1 and G2 phase) where cells mainly grow and prepare the subsequent cell division. Gap phases enclose the synthesis phase (S phase) in which DNA is duplicated meaning that 1-chromatid-chromosomes (chromatids, 2C) become 2-chromatid-chromosomes (sister chromatids, 4C). The nuclear and the cytoplasmic division takes place in mitosis (M phase) and cytokinesis. Cell division is asymmetric which results in a small daughter and a large mother cell [10].

The M phase itself consists of several phases: prophase, metaphase, anaphase and telophase [11]. In prophase, the nuclear envelope dissolves and the mitotic spindle forms. In metaphase, chromosomes align in the middle of the spindle and form the metaphase plate. In anaphase, sister chromatids separate and the spindle pulls chromatids to the opposite sides in the cell. In telophase, nuclear envelopes form around each complete set of chromosomes.

The cell cycle is a highly controlled process with a number of check points. The checkpoint START regulates the G1 to S phase transition. At this point, the decision to enter a new cell cycle is made [8]. The G2-M control point will be passed if DNA duplication was successful. The checkpoint EXIT releases cells from M into G1 phase if the nuclear and the cyctoplasmic division was correctly completed. Check point transitions are irreversible and, therefore, provide directionality to the cell cycle [11, 12]. Cell cycle deregulation leads to uncontrolled cell proliferation as well as genomic and chromosomal instability which are main causes for cancer in multicellular eukaryotic organisms [13, 14, 15, 16].

Cyclins, cyclin dependent kinases (CDK) and cyclin dependent kinase inhibitors (CKI) are important molecules which are involved in cell cycle regulation [8]. The main kinase in yeast is Cdc28. Cdc28 binds its corresponding cyclins to get activated [17]. The G1 cyclin Cln2, the B-type cyclin Clb5 and the CKI Sic1 are key regulators of the G1 to S phase transition [18]. Their

[3] ensembl release 94 - October 2018: `http://ensemblgenomes.org/`

[4] yourgenome - copyright information [2016-07-07]: `https://www.yourgenome.org`

expression peaks are close to the transition point [19, 20, 21]. Sic1 inhibits active Clb5-Cdc28. Sic1 phosphorylation by active Cln2-Cdc28 and active Clb5-Cdc28 leads to its destruction and, finally, to S phase transition [22]. Active Cln2-Cdc28 cause bud formation.

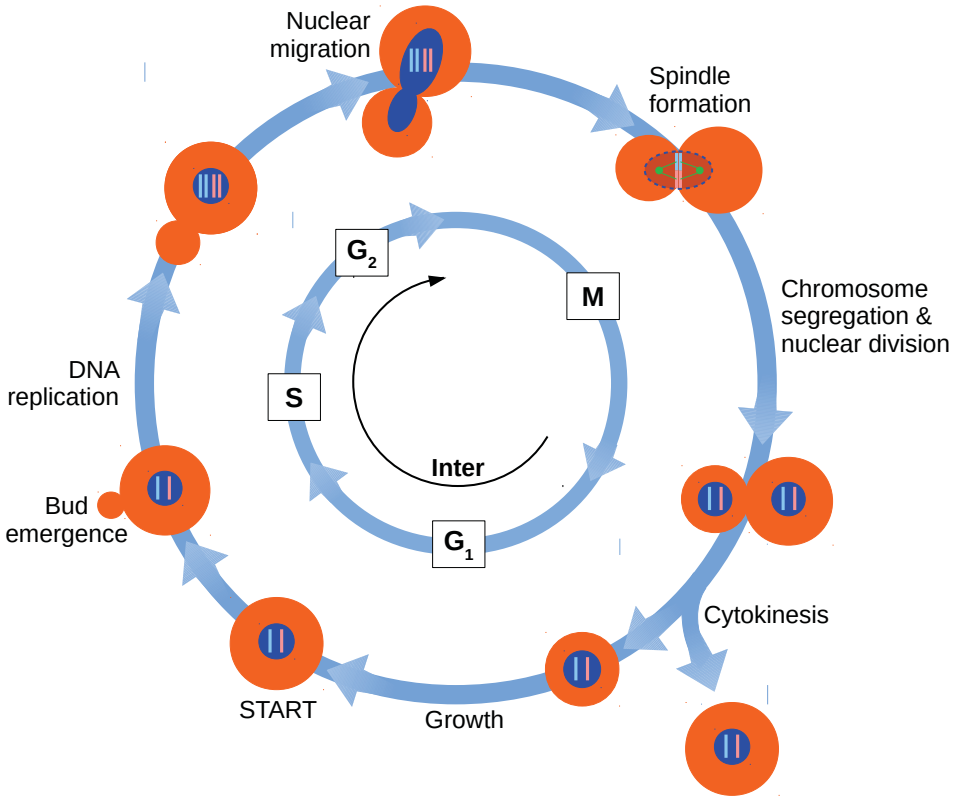

Figure 1.1: Mitotic cell cycle. The mitotic cell cycle of diploid cells (2N) is composed of an interphase and mitosis. The interphase consists of two gap phases (G1 and G2 phase) which enclose the synthesis phase (S phase). In the gap phases, cells mainly grow and prepare the subsequent cell division. In S phase, cells duplicate DNA and cells with two chromatids (2C) become cells with two sister chromatids (4C). The nuclear and the cytoplasmic division take place in several phases during mitosis (M phase) and cytokinesis. START is a control point which regulates the G1 to S phase transition. In haploid cells (1N), cell cycle stages are equivalent but the number of chromosomes is halved (1C and 2C).

1.3 Synchronizing cell populations

In unsynchronized cell populations, each cell has its own timing. Consequently, individual cells are in different cell cycle phases and have different ages [23]. The life span of a cell is about 20-30 cell divisions [24, 25]. The whole population represents a mixture of different cells. Typically, molecular compositions change over the cell cycle and individual molecules are present in specific cell cycle phases [26, 27]. It is often not possible to measure these molecules in heterogenous cell populations. Synchronization methods are used to accumulate cells in a specific cell cycle phase. As a result, every cell in the cell population starts in the same cell cycle phase after release from the synchronization procedure. In this way, molecular numbers are accumulated as well and become measurable. Cell cycle synchronization does not persist forever. Cells start to desynchronize immediately after release [28, 29]. For this reason, “synchronization” is actually a phase resetting because there is no stable phase relation between cells (phase-locking) [30].

Different synchronization methods synchronize cells in different cell cycle phases. Chemical synchronizations, so called “block- and release synchronizations”, are stronger than physical synchronizations. They block cell division but cell growth and protein synthesis continue which can cause artifacts [28]. Synchronization by α-factor is the most frequently used chemical synchronization method. MATα cells produce mating pheromone α-factor to block MATa cells in G1 phase by inhibiting active Cln2-Cdc28 [31]. Cells quickly enter into S phase after

release. Further, cells can be synchronized by hydroxyurea [32, 33] and nocodazole [32, 34]. Hydroxyurea inhibits the ribunucleotide reductase and leads to S phase arrest. The ribunucleotide reductase is involved in the formation of deoxyribonucleotides which are consumed in DNA synthesis. Cells go through S phase after release. Differently, nocodazole inhibits the microtubule polymerization and causes a G2 arrest. Cells have a shorter G2 phase.

Centrifugal elutriation is a physical synchronization method [28, 32]. Cells are sorted by size, mass and shape using centrifugal force and counterflowing media. Thus, small G1 daughter cells with prolonged G1 phases are selected. In contrast to chemical synchronization methods, cells are not perturbed in the coordination of cell division but their synchronization is less efficient [28, 32, 35]. However, the centrifugal force can generate a stress response. Beside chemical and physical synchronizations, genetic synchronization methods are used as well [36, 35]. An example are temperature sensitive cdc14 or cdc15 mutants which synchronize due to changes in temperature [37]. In general, synchronization efficiency vary in different yeast strains [35].

1.4 Measuring single cell and population data

Experimental techniques have to be adapted to the considered model organism, specific biological questions and technical means. Measurements can be done in single cells or in cell populations. Population data represent population averages and cell-to-cell variability is not covered [38]. If population data are used, it is assumed that cell populations have a single cellular state and deviations from that state have no functional significance. Therefore, cell populations are assumed to be homogenous and it is sufficient to represent them by the mean. Consequently, it is not possible to extract informations about individual cells or about the existence of subpopulations.

Single cell data reveal heterogeneity of cell populations and enable the analysis of cell-to-cell variability, multicellular states and the functional significance of noise [38]. Multicellular states result in cell populations which are composed of subpopulations. In this way, cells causing diseases can be identified among the outliers [39]. Single cell data have been found especially insightful for gene expression studies [40]. A main finding was that stochasticity in gene expression is generated by intrinsic and extrinsic noise [41, 42, 43, 44]. Intrinsic noise results from the randomness of the biochemical reaction itself and extrinsic noise comes from fluctuations in other cellular components. It was found that gene expression in bacteria is dominated by intrinsic noise. In contrast, it was found that gene expression noise in yeast is primarily extrinsic and causes correlations between fluctuating genes [45]. It was also shown that extrinsic noise in yeast is mostly influenced by cell size and shape [46].

There is a number of experimental techniques applicable in single cells and in cell populations as genome sequencing [3, 47], RNA sequencing [48, 49] and Western blotting [50, 51]. In recent times, scientists spend a lot of effort to develop advanced experimental techniques on the single cell level. Further examples are single molecule RNA *in situ* hybridization (smFISH) [52], gene expression profiling by flow cytometry [53, 54], gene expression profiling by fluorescent *in situ* sequencing (FISSEQ) [55] and time lapse microscopy in combination with cell tracking [56].

1.5 Mathematical modeling in systems biology

Systems biology combines experimental studies with mathematical modeling [8]. Mathematical models of biological processes allow for testing biological hypothesis and making quantitative predictions. Therefore, mathematical modeling is mandatory to understand underlying mechanism, especially if experimental evidences are still missing. The following citation which is taken from [57] perfectly describes what systems biology is about:

> *"[...] Because a system is not just an assembly of genes and proteins, its properties cannot be fully understood merely by drawing diagrams of their interconnections. Although such a diagram represents an important first step, it is analogous to a static roadmap, whereas what we really seek to know are the traffic patterns, why such traffic patterns emerge, and how we can control them. [...]"*

Systems biology is about studying functional connections between molecular components and not only about finding these components.

A mathematical model can be deterministic or stochastic and has to be adjusted by experimental data to make useful model predictions [58, 59]. In contrast to physics, only a few model parameters can be measured directly and most remain unknown [59, 60]. Technical limitations are the main reason why model parameters cannot be experimentally determined [60]. If parameters cannot be measured, we have to estimate them [61]. Parameters can be estimated from measurements which represent molecular components included in the mathematical model.

Availability of experimental data is crucial in mathematical modeling and still limited in most cases. Some data are published or available in databases and can easily be used to estimate your model parameters and, in turn, to answer your own biological questions. It is more common that available data are not usable regarding your specific modeling approach or that required data are not measured. Moreover, not every molecular component is measurable at all. Furthermore, every data type has its own characteristic which cannot always be combined with other data types in a straightforward way.

Several software packages for parameter estimation are available. Examples are D2D [62], dMod [63], AMIGO2 [64], MEIGO [65], PESTO [66], COPASI [67] and SBML-PET [68]. These tools are generally developed to estimate parameters of deterministic model systems described by ordinary differential equations (ODE). ODE systems are most frequently used in systems biology to study molecular interactions.

2. General interest of this study

Biological questions

In most studies, people analyze synchronized cell populations to finally draw conclusions on unsynchronized cells. It is known that synchronization affects not only the cell cycle behavior but also the cell morphology and the overall functioning of the cell [28, 35, 69]. In some cases, there is an experimental need to work with synchronized cell populations, e.g. to measure regulation of cell proliferation, gene expression or metabolic pathways [35]. However, synchronization effects cannot be ignored in data analysis or in conclusions about biological processes.

Synchronization can be interpreted as cellular stress where cells will respond to. It is of interest to know how cells respond to those synchronization specific stresses. Therefore, we ask the following questions:

1. How does synchronization affects the cell cycle?
2. How does different synchronization methods influence the cell cycle?
3. How different are unsynchronized and synchronized cells?
4. Is it even possible to derive general cell cycle behavior from synchronized cell populations?
5. What did we really learn the last years about the cell cycle?

In this study, we use phase-resolved mRNA distributions of mRNA species *SIC1*, *CLN2* and *CLB5* which are measured by smFISH. These data were originally recorded to analyze transcriptional timing and noise [52]. The important feature for this study is that mRNA numbers were counted in unsynchronized cells. To analyze the cell cycle behavior in synchronized cells, we use protein time courses of protein species Sic1, Cln2 and Clb5 which are measured by Western blotting. We decided for synchronization by centrifugal elutriation, α-factor, hydroxyurea and nocodazole. Western blot data give relative protein numbers. We use absolute protein numbers per cell reported by the PaxDb database[5] to normalize them.

Mathematical modeling tasks

Data availability is the driving force for mathematical modeling tasks. We have data sets of different quality, quantity and conditions. mRNA numbers are absolute and given as phase-resolved distributions whereas protein numbers are relative and given as time courses. The number of technical and biological replicates differ in both measurements. Additionally, the number of cells measured per mRNA species and cell cycle phase differ in mRNA measurements. Moreover, we consider unsynchronized single cell data on the mRNA level and synchronized cell population data on the protein level.

Phase-resolved mRNA distributions are the main cause for choosing a stochastic model system described by the chemical master equation (CME). Stochastic models are less frequently parameterized in systems biology compared to deterministic models. The first task follows from the stochastic modeling approach:

1. How to parameterize a stochastic model?

In the optimization problem of this study, we have to integrate phase-resolved mRNA distributions and protein time courses in a common parameter estimation. Combining different data types is not popular in the field of parameter estimation. Even if people estimate parameters of stochastic models, they typically use one specific data type. Examples are given in [70] where parameters of the CME are estimated from smFISH data and in [71] where parameters of nonlinear stochastic differential equations are estimated from noisy time series data. At this point, the second task arises:

[5] freely available on `https://pax-db.org/`

2. How to estimate parameters from different data types?

In this study, we parameterize a stochastic model of the cell cycle by using optimization methods developed for deterministic model systems. We use a 2-step optimization method to combine phase-resolved mRNA distributions and protein time courses. In the first step, mRNA parameters are estimated from smFISH data. In the second step, mRNA parameters are re-estimated and protein parameters estimated from Western blot data. In this way, we can predict synchronization specific mRNA dynamics.

Vision

Modelers are often in search of perfect data. Perfect data have to be less noisy and need to be time series data. Furthermore, absolute measurements are preferred over relative measurements and have to be available for every desired model component. Repeated measurements are mandatory and different experimental conditions are required. Obviously, perfect data are far from reality and we have to deal with available data. Data availability is restricted to experimental techniques and costs. Costs include the financial value and manpower.

What we really need to save resources and in our case to estimate parameters more efficiently is improvement in experimental techniques and in mathematical modeling. Experimental techniques are often limited to specific organisms and not generally usable. Additionally, parameter estimation methods are restricted to specific model and data types. In general, we should always ask ourselves the following questions:

1. What can we learn from existing data?
2. Is there really a need for new measurements?

II Methods

3 Experimental data 23
3.1 Counting mRNA molecules in single cells by smFISH
3.2 Measuring relative protein abundances of cell populations by Western blotting
3.3 Pre-processing Western blot data
3.4 Normalizing Western blot data by using absolute protein numbers from the PaxDb database

4 Mathematical modeling 27
4.1 Modeling the chemical reaction system
4.2 Formulating the chemical master equation
4.3 Transition to the reaction rate equation
4.4 Using Gillespie's stochastic simulation algorithm
4.5 Representation of the experimental data by the stochastic model

5 Parameter estimation 45
5.1 Estimating parameters in ODE systems
5.2 Applying the maximum likelihood approach
5.3 Combining count and time series data in a 2-step-optimization
5.4 Performing global optimization by multi-start local optimization

6 Identifiability analysis 51
6.1 Introducing the concept of parameter identifiability
6.2 Working with profile likelihoods
6.3 Classification of protein profile effects

7 L_1 regularization 57
7.1 Regularizing the likelihood function
7.2 Identifying protein parameters not covered by the data
7.3 Calculating fold changes between mRNA parameters of unsynchronized and synchronized cells

3. Experimental data

3.1 Counting mRNA molecules in single cells by smFISH

The absolute number of gene transcripts in unsynchronized single cells can be measured by smFISH [72, 73, 74]. In this method, fluorescently labeled DNA probes hybridize with the target mRNA sequences and, therefore, become detectable in fluorescence microscopy. We measured mRNA numbers for key regulators of the G1 to S phase transition: *SIC1*, *CLN2* and *CLB5*. In total, we combined measurements of more than 900 cells for each mRNA species by pooling up to four biological replicates. We used the haploid yeast strain BY4741 (MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0).

We counted the absolute number of mRNA molecules in each cell and assigned respective cells to a specific cell cycle phase by using morphological markers. Markers are presence and size of a bud, morphology of the nucleus, number and localization of spindle pole bodies and localization of Whi5. Whi5 is a transcriptional repressor which is located in the nucleus between late M and early G1 phase and in the cytoplasm in remaining cell cycle phases [75]. We distinguish between seven cell cycle phases: early G1, late G1, S, G2, pro-/metaphase (P/M), anaphase (Ana) and telophase/cytokinesis (T/C). Thus, we got mRNA distributions per cell cycle phase. The number of cells measured in each phase is directly proportional to the cell cycle phase length and dependent on the cell division time [72]. A detailed description of the data is given in [52].

3.2 Measuring relative protein abundances of cell populations by Western blotting

The relative number of protein molecules of synchronized cell populations can be quantified by Western blotting. We used centrifugal elutriation, α-factor, hydroxyurea and nocodazole to synchronize cells. Synchronization efficiency was determined by using the bud index (# buds/# cells) and DNA content (1N and 1C or 1N and 2C). In synchronized cell populations, buds arise simultaneously and the bud index is high. Similarly, the whole population have the same DNA content.

We measured protein products of mRNA species mentioned in the previous section in the same haploid yeast strain: Sic1, Cln2 and Clb5. In order to interpret Western blot band intensities as a measure for protein expression differences, we tagged low abundant target proteins with the same 3xFlag-tag. For quantitative Western blotting, the protein samples were separated via gel electrophoresis and transferred to a blot membrane. The tagged proteins were visualized by incubating the membrane with mouse anti-Flag antibodies and a secondary anti mouse antibody with a fluorescent label. For normalization purposes, a rabbit anti-glucose-6P-DH antibody against a housekeeping protein was used and visualized with a fluorescent marker which is detectable in a second channel.

These measurements give time courses over at least one cell division with a sampling rate of ten minutes. We used the number of cell counts and the cell volume to determine the cell division time. The number of cell counts is constant until cells divide and doubled after a complete cell division. Cell volume increases over the cell cycle and stagnates after cell division. Determined cell division times are only approximately true due to inaccuracies resulting from a rather large sampling rate. Since we expect most synchronization effects in the first cell cycle after release, we only consider the first cycle in following calculations.

We normalized Western blot data by multiplying band intensities of the respective protein with ratios of total and time dependent band intensities of the housekeeping protein ("total protein normalization"). In each experiment, we applied two samples for every time point on the gel. We treat these samples as technical replicates. Up to three gels are required for a complete

measurement. With the exception of elutriation, the first cycle is measured on the same gel. Even if we have a second experiment for any synchronization method, it is still a technical replicate.

3.3 Pre-processing Western blot data

Assuming identically and independently distributed (iid) data points which are sampled from the same probability density function with mean μ and variance σ^2, the sum over all data points follows for $N \to \infty$ a normal distribution with mean $N\mu$ and variance σ^2/N according to the central limit theorem (CLT) [61, 76]. The standard deviation $\sigma/\sqrt{N}$ is named standard error of the mean (SEM). The SEM falls with $1/\sqrt{N}$ and shows that we need four times more data points to get a twice as good resolution.

Repeated time course measurements relate to the CLT. There are two possibilities to determine the measurement error σ. Either errors are determined simultaneously with the optimization problem or data are pre-processed [77]. We decided for data pre-processing by using the R function `alignME()` of the R package **blotIt2**[6] which is designed to align immunoblots [78]. The alignment is needed to make different measurements comparable. Differences consist in gel quality, protein concentration per lysate, blotting efficiency or intensity. One difference between error estimation during pre-processing and as part of the optimization problem is that for pre-processing one needs more than one replicate. Error estimation as part of the optimization problem is possible already with individual replicates.

We estimated measurement specific scaling factors and a constant relative error to determine values and errors over one cell division for every synchronization method (see Appendix D.1 for calculation details). It was shown that biological variability and measurement noise are multiplicative and log-normally distributed in immunoblotting experiments [77, 79, 80]. However, most statistical methods are based on additive normally distributed noise and, therefore, a log-transformation of the data is favorable. Unfortunately, Cln2 time courses of elutriated and hydroxyurea synchronized cells show negative band intensities for a few time points. There is at least one non-negative band intensity for relevant time points. We set negative band intensities to one because pre-processing did not work error-free for negative or zero band intensities. These values become close to zero after pre-processing. For technical reasons we decided to use non-transformed data. The logarithm is defined for $\mathbb{R}^+ \setminus \{0\}$ and $log(1) = 0$ where pre-processing is still not working correctly. Nevertheless, α-factor and nocodazole synchronized cells show approximately the same pre-processed values and errors for log- and non-transformed data.

3.4 Normalizing Western blot data by using absolute protein numbers from the PaxDb database

Pre-processed Western blot data are still relative. We used absolut protein numbers reported by the PaxDb database to transform relative to absolute protein numbers. The PaxDb database combines genome-wide proteome quantifications which range from affinity-based and biophysical methods to the large array of mass spectrometry-based quantification techniques [81]. In the integrated data set of budding yeast, the weighted average of 17 quantitative proteomics data sets are reported.

Normalization was done by first multiplying the absolut protein number per cell given by the PaxDb database in parts per million (ppm) with the total number of proteins per cell ($\approx 45 \times 10^6$) reported in [82] divided by 10^6 (see Table 3.1). Secondly, we normalized the mean value of the pre-processed protein time course to the calculated absolute protein number by multiplying with the ratio of the absolute protein number and the mean value of the pre-processed protein time course. Errors of pre-processed data are normalized by multiplying with the same ratio.

[6] freely available on `http://www.fdmold.uni-freiburg.de/~kaschek/`

Protein	PaxDb [ppm]	PaxDb×45 [#]
Sic1	9.47	426.15
Cln2	13.60	612.00
Clb5	4.59	206.55

Table 3.1: Absolute protein numbers. This table represents absolut protein numbers given by the PaxDb database in parts per million (second column) [81] and transformed values which result from the multiplication with the total number of proteins reported in (third column) [82]. The total number of proteins is calculated by $(45 \times 10^6)/(1 \times 10^6) = 45$.

4. Mathematical modeling

All definitions in this chapter are taken from [83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93]. Different references are indicated.

4.1 Modeling the chemical reaction system

Our mathematical model represents a small chemical reaction system which includes key regulators of the G1 to S phase transition, inspired by [50]. Summarizing, we look at $N = 9$ species (see Figure 4.1 and Table 4.1). Cdc28 is assumed to be unchanged and available in sufficient quantity over the cell cycle [5]. Therefore, Cdc28 is not explicitly modeled. The chemical reaction system separates into two parts: mRNA and protein level. On the mRNA level, gene transcripts are independently produced and degraded. Protein productions, degradations and interactions are coupled to the mRNA level or other protein reactions. In total, we regard $M = 17$ reactions (see Table 4.2). We only consider elementary chemical reactions for the reliability of the CME and Gillespie's stochastic simulation algorithm (SSA) [94].

Elementary chemical reactions describe individual reactions which are required to go from initial reactant molecules to final products [90]. For example, a reversible reaction S $\rightleftarrows$ P divides into two reactions S $\longrightarrow$ P and P $\longrightarrow$ S. The molecularity of a chemical reaction is given by the number of reactant molecules entering the reaction [90]. We consider uni- and bimolecular reactions which represent linear and quadratic birth-death processes. Reactions of higher molecularities do not count for elementary chemical reactions because they decompose into sequences of uni- and bimolecular reactions. For instance, a trimolecular reaction divides into a bi- and an unimolecular reaction [88, 90]. Regarding deterministic chemical reaction systems, elementary chemical reactions translate into mass action kinetics.

We assume that genes are differently expressed in high and low transcription regions which are triggered by signals $s_1(t)$ to $s_3(t)$. If signals turn on, genes are transcribed with stochastic rate constants $c_{j,high}$ and otherwise with $c_{j,low}$. In practice, signal dependent transcriptions are defined as time dependent stochastic rates $c_j(t)$ where high transcription regions are defined by transcription start (e.g. $t_{s_{1,0}}$) and end (e.g. $t_{s_{1,e}}$) times (see Table 4.2). mRNA species *CLN2* and *CLB5* are modeled with two high transcription regions. Their start and end times have additional subscripts for the the first ($1st$) and the second ($2nd$) high transcription region.

In general, chemical reaction systems are described by a continuous deterministic or a discrete stochastic formulation. Both formalisms are based on the same assumption of a well-stirred system with a constant system volume V in thermal, not chemical, equilibrium at a constant absolut temperature T. This assumption allows to describe the system state by molecular populations only and to ignore positions and velocities of the individual molecules. That means that molecular positions are considered as random variables that are uniformly distributed throughout the system volume V. Molecular velocities are considered as random variables that are normally distributed with a zero valued mean and variance k_BT/m leading to the Maxwell-Boltzmann distribution

$$p_{MB}(\boldsymbol{v}) = \left(\frac{m}{2\pi k_B T}\right)^{3/2} exp\left(-\frac{m|\boldsymbol{v}|^2}{2k_B T}\right) \tag{4.1}$$

with m the molecular mass, k_B the Boltzmann constant and $\boldsymbol{v} = (v_x, v_y, v_z)$ the velocity vector in Cartesian coordinates. In addition, we expect validity of this assumption for any constant-temperature dilute-gas system in which nonreactive molecular collisions occur more frequently than reactive molecular collisions which result in a populational change.

In the stochastic modeling approach, we consider a small system volume with a few number of reactant molecules (microscopic view). These processes are not fully predictable and typically governed by a jump type Markov process with continuous time and a discrete system state

[88, 95]. Starting several times from the same initial system state leads to different final system states. This behavior comes from specifying molecular populations regardless of molecular positions and velocities. The chemical reaction system is described by the CME and includes stochastic reaction constants c_j. That is the deterministic time evolution of the system's probability to be in a particular state at time t and consists of a set of ODEs, one ODE for each possible system state. In our system, we get 100^9 possible system states for molecular numbers ranging from 0 to 99 for each species.

In comparison, we consider a large system volume with a large number of reactant molecules in the deterministic modeling approach (macroscopic view). These processes are predictable. Starting from the same initial system state always leads to the same final system state. The chemical reaction system is described by a set of coupled ODEs, called reaction rate equation (RRE), with one differential equation per molecular species. Our system end up with nine ODEs. Here, reaction constants k_j are deterministic and both system state and time are continuous.

Fluctuations and correlations are only considered in the stochastic modeling approach. The stochastic formulation is valid whenever the deterministic formulation is valid but not the other way around [84]. In the thermodynamic limit, the stochastic formulation reduces to the deterministic formulation (see Section 4.3). This is why both formulations are based on the same assumption. Analytical solutions can be calculated for some simple examples. There are fewer examples for the CME compared to the RRE. That is why numerical solutions are used. The CME is exact, but not useful for numerical calculations. Typically, the SSA is used which is exact as well. In contrast, the infinitesimal time increment dt in the numerical solution of an ODE is approximated by a finite time steps Δt and not exact.

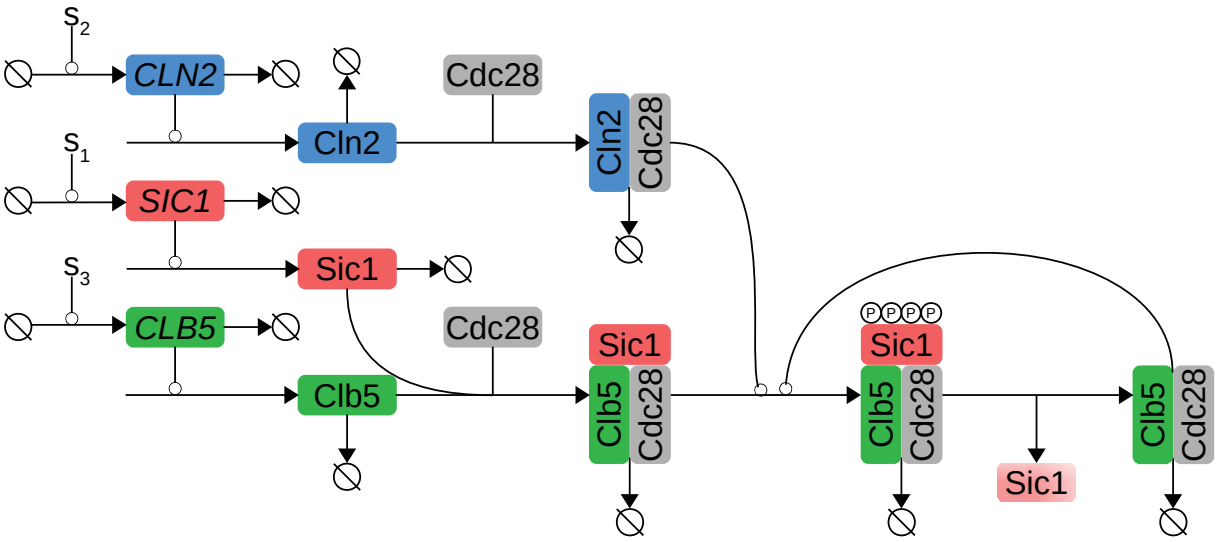

Figure 4.1: Chemical reaction system. The chemical reaction system describes productions and degradations of mRNA and protein species as well as protein interactions for key regulators of the G1 to S phase transition. Signals $s_1(t)$ to $s_3(t)$ induce high and low transcription regions and are marked as s_1 to s_3. *CLN2*, *SIC1* and *CLB5* indicate mRNA species. Cln2, Sic1 and Clb5 indicate protein species. Edges with arrowheads represent reactions whereas edges with open circles show activations. mRNA production leads to protein production and subsequently to the formation of active cyclin-Cdc28. Sic1 inhibits active Clb5-Cdc28. Active Cln2-Cdc28 and active Clb5-Cdc28 phosphorylate Sic1 and release active Clb5-Cdc28 from Sic1 to induce S phase transition [22]. Phosphorylated Sic1 gets destructed. Active Cln2-Cdc28 cause bud formation. Cdc28 is assumed to be unchanged and available in sufficient quantity over the cell cycle [5]. Therefore, Cdc28 is not explicitly modeled.

4.2 Formulating the chemical master equation

We assume a chemical reaction system with constant system volume V and X_i molecules of N species S_i (see Table 4.1). The system state vector $\boldsymbol{X}(t) = (X_1(t), ..., X_N(t))$ is a random integer variable where $X_i(t)$ describes the number of molecules X_i at time t. The number of molecules

X_i changes by M elementary chemical reactions R_j which are characterized by stochastic rate constants c_j (see Table 4.2). We consider unidirectional elementary chemical reactions of the following type:

$$\varnothing \longrightarrow S_i, \tag{4.2a}$$
$$S_h \longrightarrow \varnothing, \tag{4.2b}$$
$$S_h \longrightarrow S_i \qquad (h \neq i), \tag{4.2c}$$
$$S_g + S_h \longrightarrow S_i \qquad (g \neq h, h \neq i, g \neq i), \tag{4.2d}$$

where S_g and S_h are reaction educts and S_i is the reaction product [95].

Unimolecular reactions include conversion and outflow reactions (Equations (4.2c) and (4.2b)) and are independent of the system volume V. Inflow and bimolecular reactions (Equation (4.2a) and (4.2d)) are dependent of the system volume V. Inflow reactions are directly proportional to the system volume V and bimolecular reactions are inversely proportional to it (see Figure 4.4). Consequently, a collision-initiated conversion of two reactant molecules takes more time in a large system volume V compared to a small one. State-change vector $\boldsymbol{\nu}_j = (\nu_{1j}, ..., \nu_{Nj})$ is the j-th column of the stoichiometric matrix $\boldsymbol{\nu}$ (see Table F.1) where each integer value is defined as

> $\nu_{ij} \triangleq$ the change in species S_i caused by one reaction R_j.

Thus, system state $\boldsymbol{X}(t)$ jumps to $\boldsymbol{X}(t) + \boldsymbol{\nu}_j$ after one reaction R_j.

The fundamental hypothesis of a stochastic chemical reaction system is the stochastic rate constant c_j defined as

> $c_j dt \triangleq$ the probability that a particular combination of R_j reactant molecules will react accordingly in the next infinitesimal time interval $[t, t + dt)$.

The propensity function $a_j(\boldsymbol{X}(t))$ emerges from the fundamental hypothesis and is defined as

> $a_j(\boldsymbol{X}(t))dt \triangleq$ the probability that one reaction R_j will occur somewhere inside the system volume V in the next infinitesimal time interval $[t, t + dt)$, given the system is in state $\boldsymbol{X}(t)$.

The value of the propensity function $a_j(\boldsymbol{X}(t))$ is the product of the stochastic rate constant c_j and the combinatorial function $h_j(\boldsymbol{X}(t))$ which is defined as

> $h_j(\boldsymbol{X}(t)) \triangleq$ the total number of distinct combinations of R_j reactant molecules, given the system state $\boldsymbol{X}(t)$.

For the reaction types defined in Equation (4.2) the following combinatorial functions arise:

$$h_j(\boldsymbol{X}(t)) = 1, \tag{4.3a}$$
$$h_j(\boldsymbol{X}(t)) = X_h(t), \tag{4.3b}$$
$$h_j(\boldsymbol{X}(t)) = X_h(t), \tag{4.3c}$$
$$h_j(\boldsymbol{X}(t)) = X_h(t)X_g(t), \tag{4.3d}$$

and, therefore, the propensity functions are:

$$a_j(\boldsymbol{X}(t)) = c_j, \tag{4.4a}$$
$$a_j(\boldsymbol{X}(t)) = c_j X_h(t), \tag{4.4b}$$
$$a_j(\boldsymbol{X}(t)) = c_j X_h(t), \tag{4.4c}$$
$$a_j(\boldsymbol{X}(t)) = c_j X_h(t)X_g(t). \tag{4.4d}$$

The propensities of our chemical reaction system are given in Table 4.2.

The CME is based on the conditional probability

$P(\boldsymbol{x}, t|\boldsymbol{x}_0, t_0) \triangleq$ the probability that there will be $\boldsymbol{X}(t) = \boldsymbol{x}$ molecules in the system volume V at time t, given the system was in state $\boldsymbol{X}_0(t_0) = \boldsymbol{x}_0$ before.

Small letters $\boldsymbol{x} = (x_1, ..., x_N)$ indicate integer values the system state $\boldsymbol{X}(t)$ can achieve. To derive the CME we start from the initial state $\boldsymbol{X}_0(t_0) = \boldsymbol{x}_0$ with probability $P(\boldsymbol{x}_0, t_0)$ and ask for all possible paths through the intermediate system states $\boldsymbol{Z}_j(t) = \boldsymbol{z}_j$ with probabilities $P(\boldsymbol{z}_j, t|\boldsymbol{x}_0, t_0)$ to reach the final system state $\boldsymbol{X}(t + dt) = \boldsymbol{x}$ with probability $P(\boldsymbol{x}, t + dt|\boldsymbol{z}_j, t; \boldsymbol{x}_0, t_0)$ (see Figure 4.2). The joint probability $P(\boldsymbol{x}, t + dt; \boldsymbol{z}_j, t; \boldsymbol{x}_0, t_0)$ of a whole j-th path is given by

$$P(\boldsymbol{x}, t + dt; \boldsymbol{z}_j, t; \boldsymbol{x}_0, t_0) = P(\boldsymbol{x}, t + dt|\boldsymbol{z}_j, t; \boldsymbol{x}_0, t_0)P(\boldsymbol{z}_j, t|\boldsymbol{x}_0, t_0)P(\boldsymbol{x}_0, t_0). \tag{4.5}$$

Markov processes are characterized by regarding the most recent value of the process instead of evaluating the whole process history. Therefore, Equation (4.5) becomes

$$P(\boldsymbol{x}, t + dt; \boldsymbol{z}_j, t; \boldsymbol{x}_0, t_0) = P(\boldsymbol{x}, t + dt|\boldsymbol{z}_j, t)P(\boldsymbol{z}_j, t|\boldsymbol{x}_0, t_0)P(\boldsymbol{x}_0, t_0) \tag{4.6}$$

where $P(\boldsymbol{x}, t + dt|\boldsymbol{z}_j, t)$ and $P(\boldsymbol{z}_j, t|\boldsymbol{x}_0, t_0)$ are so called transition probabilities. For calculating the joint probability $P(\boldsymbol{x}, t + dt; \boldsymbol{x}_0, t_0)$, we sum over all possible intermediate system states $\boldsymbol{z}_j$ and end up with the forward Chapman-Kolmogorov equation (CKE):

$$\begin{aligned} P(\boldsymbol{x}, t + dt; \boldsymbol{x}_0, t_0) &= P(\boldsymbol{x}, t + dt|\boldsymbol{x}_0, t_0)P(\boldsymbol{x}_0, t_0) \\ P(\boldsymbol{x}, t + dt|\boldsymbol{x}_0, t_0)P(\boldsymbol{x}_0, t_0) &= \sum_{\boldsymbol{z}_j} P(\boldsymbol{x}, t + dt|\boldsymbol{z}_j, t)P(\boldsymbol{z}_j, t|\boldsymbol{x}_0, t_0)P(\boldsymbol{x}_0, t_0) \\ P(\boldsymbol{x}, t + dt|\boldsymbol{x}_0, t_0) &= \sum_{\boldsymbol{z}_j} P(\boldsymbol{x}, t + dt|\boldsymbol{z}_j, t)P(\boldsymbol{z}_j, t|\boldsymbol{x}_0, t_0). \end{aligned} \tag{4.7}$$

There are two possibilities for the intermediate system state $\boldsymbol{z}_j$. If $\boldsymbol{z}_j = \boldsymbol{z}_0 = \boldsymbol{x}$, no reaction takes place and the probability that no reaction fires to reach $\boldsymbol{x}$ is

$$\begin{aligned} P(\boldsymbol{x}, t + dt|\boldsymbol{z}_0, t) &= P(\boldsymbol{x}, t + dt|\boldsymbol{x}, t) \\ &= 1 - a_0(\boldsymbol{x})dt \end{aligned} \tag{4.8}$$

with $a_0(\boldsymbol{x}) = \sum_{j=1}^{M} a_j(\boldsymbol{x})$ and $a_0(\boldsymbol{x})dt$ the probability that any reaction fires over $[t, t + dt)$. For $\boldsymbol{z}_j$ being one reaction apart from $\boldsymbol{x}$, reaction R_j fires. The probability that reaction R_j fires over $[t, t + dt)$ to reach $\boldsymbol{x}$ becomes

$$\begin{aligned} P(\boldsymbol{x}, t + dt|\boldsymbol{z}_j, t) &= P(\boldsymbol{x}, t + dt|\boldsymbol{x} - \boldsymbol{\nu}_j, t) \\ &= a_j(\boldsymbol{x} - \boldsymbol{\nu}_j)dt. \end{aligned} \tag{4.9}$$

Since we consider just one reaction in the time interval $[t, t + dt)$, transitions from $\boldsymbol{x_0}$ to $\boldsymbol{x}$ which are more than one reaction apart are excluded.

With $\boldsymbol{z}_0 = \boldsymbol{x}$ and $\boldsymbol{z}_j = \boldsymbol{x} - \boldsymbol{\nu}_j$ the CKE becomes

$$\begin{aligned} P(\boldsymbol{x}, t + dt|\boldsymbol{x}_0, t_0) &= \sum_{j=0}^{M} P(\boldsymbol{x}, t + dt|\boldsymbol{z}_j, t)P(\boldsymbol{z}_j, t|\boldsymbol{x}_0, t_0) \\ &= P(\boldsymbol{x}, t + dt|\boldsymbol{x}, t)P(\boldsymbol{x}, t|\boldsymbol{x}_0, t_0) \\ &+ \sum_{j=1}^{M} P(\boldsymbol{x}, t + dt|\boldsymbol{x} - \boldsymbol{\nu}_j, t)P(\boldsymbol{x} - \boldsymbol{\nu}_j, t|\boldsymbol{x}_0, t_0). \end{aligned} \tag{4.10}$$

Inserting Equations (4.8) and (4.9) into Equation (4.10) leads to

$$\begin{aligned} P(\boldsymbol{x}, t + dt|\boldsymbol{x}_0, t_0) &= \left(1 - \sum_{j=1}^{M} a_j(\boldsymbol{x})dt\right) P(\boldsymbol{x}, t|\boldsymbol{x}_0, t_0) \\ &+ \sum_{j=1}^{M} a_j(\boldsymbol{x} - \boldsymbol{\nu}_j)dtP(\boldsymbol{x} - \boldsymbol{\nu}_j, t|\boldsymbol{x}_0, t_0) \end{aligned} \tag{4.11}$$

which becomes

$$\frac{P(\boldsymbol{x}, t+dt|\boldsymbol{x}_0, t_0) - P(\boldsymbol{x}, t|\boldsymbol{x}_0, t_0)}{dt} = \sum_{j=1}^{M} [a_j(\boldsymbol{x} - \boldsymbol{\nu}_j) P(\boldsymbol{x} - \boldsymbol{\nu}_j, t|\boldsymbol{x}_0, t_0) - a_j(\boldsymbol{x}) P(\boldsymbol{x}, t|\boldsymbol{x}_0, t_0)] \tag{4.12}$$

and leads in the limit $dt \to 0$ to a special case of the differential forward CKE, known as CME:

$$d_t P(\boldsymbol{x}, t|\boldsymbol{x}_0, t_0) = \sum_{j=1}^{M} [a_j(\boldsymbol{x} - \boldsymbol{\nu}_j) P(\boldsymbol{x} - \boldsymbol{\nu}_j, t|\boldsymbol{x}_0, t_0) - a_j(\boldsymbol{x}) P(\boldsymbol{x}, t|\boldsymbol{x}_0, t_0)]. \tag{4.13}$$

In summary, the first term in the sum describes all processes leading to state $\boldsymbol{x}$ whereas the second term captures processes leaving state $\boldsymbol{x}$ (see Figure 4.3).

Molecular component	Species S_i	Number of molecules $X_i(t)$	Concentration $C_i(t)$
SIC1	S_1	$X_1(t)$	$C_1(t)$
Sic1	S_2	$X_2(t)$	$C_2(t)$
CLN2	S_3	$X_3(t)$	$C_3(t)$
Cln2	S_4	$X_4(t)$	$C_4(t)$
CLB5	S_5	$X_5(t)$	$C_5(t)$
Clb5	S_6	$X_6(t)$	$C_6(t)$
Sic1Clb5	S_7	$X_7(t)$	$C_7(t)$
P_Sic1Clb5	S_8	$X_8(t)$	$C_8(t)$
$\text{Clb5}_{\text{active}}$	S_9	$X_9(t)$	$C_9(t)$

Table 4.1: Components of the chemical reaction system. This table shows $N = 9$ molecular components integrated in our chemical reaction system (see Figure 4.1). We assign these components to species S_i, number of molecules X_i and concentrations C_i for $i = 1, ..., N$.

4.3 Transition to the reaction rate equation

Now, we introduce a leap time $\tilde{\tau} > 0$. The first request for $\tilde{\tau}$ is to be small enough to satisfy the leap condition that in the time interval $[t, t+\tilde{\tau})$ relatively few reactions R_j take place and propensities $a_j(\boldsymbol{X}(t))$ do not change significantly. Under this assumption, the number of times reaction R_j fires in $[t, t+\tilde{\tau})$ is a pure counting process and becomes a Poisson random variable with mean and variance $a_j(\boldsymbol{X}(t))\tilde{\tau}$. It follows the tau-leaping formula:

$$\boldsymbol{X}(t+\tilde{\tau}) = \boldsymbol{X}(t) + \sum_{j=1}^{M} \mathcal{P}_j(a_j(\boldsymbol{X}(t))\tilde{\tau})\boldsymbol{\nu}_j \tag{4.14}$$

where $\mathcal{P}_j(a_j(\boldsymbol{X}(t))\tilde{\tau})$ is the probability that reaction R_j will fire $0, 1, 2, ...$ times in the time interval $[t, t+\tilde{\tau})$.

Secondly, we require $a_j(\boldsymbol{X}(t))\tilde{\tau} \gg 1$ for all reactions $1 \leq j \leq M$. Thus, $\tilde{\tau}$ has to be large enough that every reaction R_j takes place many more times than once. A Possion distribution $\mathcal{P}(\lambda)$ with a large mean value ($\lambda \gg 1$) can be approximated with a normal distribution $\mathcal{N}(\mu, \sigma^2)$ with mean μ and variance σ^2 equal to λ (see Figure E.2). In this case, the discrete random variable $\boldsymbol{X}(t)$ becomes a continuous random variable $\boldsymbol{C}(t)$ and we get

$$\boldsymbol{C}(t+\tilde{\tau}) = \boldsymbol{C}(t) + \sum_{j=1}^{M} \mathcal{N}_j(a_j(\boldsymbol{C}(t))\tilde{\tau}, a_j(\boldsymbol{C}(t))\tilde{\tau})\boldsymbol{\nu}_j \tag{4.15}$$

Reaction R_j	Reaction scheme	Propensity function $a_j(\boldsymbol{X}((t)))$	Deterministic rate constant
R_1	$\varnothing \longrightarrow S_1$	$c_1(t) = \begin{cases} c_{1,high}, & t_{s_{1,0}} < t < t_{s_{1,e}} \\ c_{1,low}, & otherwise \end{cases}$	$k_1(t)\Omega$
R_2	$S_1 \longrightarrow \varnothing$	$c_2 \cdot X_1(t)$	k_2
R_3	$S_1 \longrightarrow S_2$	$c_3 \cdot X_1(t)$	k_3
R_4	$S_2 \longrightarrow \varnothing$	$c_4 \cdot X_2(t)$	k_4
R_5	$\varnothing \longrightarrow S_3$	$c_5(t) = \begin{cases} c_{5,high}, & t_{s_{2,0,1st}} < t < t_{s_{2,e,1st}} \\ & t_{s_{2,0,2nd}} < t < t_{s_{2,e,2nd}} \\ c_{5,low}, & otherwise \end{cases}$	$k_5(t)\Omega$
R_6	$S_3 \longrightarrow \varnothing$	$c_6 \cdot X_3(t)$	k_6
R_7	$S_3 \longrightarrow S_4$	$c_7 \cdot X_3(t)$	k_7
R_8	$S_4 \longrightarrow \varnothing$	$c_8 \cdot X_4(t)$	k_8
R_9	$\varnothing \longrightarrow S_5$	$c_9(t) = \begin{cases} c_{9,high}, & t_{s_{3,0,1st}} < t < t_{s_{3,e,1st}} \\ & t_{s_{3,0,2nd}} < t < t_{s_{3,e,2nd}} \\ c_{9,low}, & otherwise \end{cases}$	$k_9(t)\Omega$
R_{10}	$S_5 \longrightarrow \varnothing$	$c_{10} \cdot X_5(t)$	k_{10}
R_{11}	$S_5 \longrightarrow S_6$	$c_{11} \cdot X_5(t)$	k_{11}
R_{12}	$S_6 \longrightarrow \varnothing$	$c_{12} \cdot X_6(t)$	k_{12}
R_{13}	$S_2 + S_6 \longrightarrow S_7$	$c_{13} \cdot X_2(t) \cdot X_6(t)$	k_{13}/Ω
R_{14}	$S_4 + S_7 \longrightarrow S_8$	$c_{14} \cdot X_2(t) \cdot X_7(t)$	k_{14}/Ω
R_{15}	$S_8 \longrightarrow S_9$	$c_{15} \cdot X_8(t)$	k_{15}
R_{16}	$S_9 \longrightarrow \varnothing$	$c_{16} \cdot X_9(t)$	k_{16}
R_{17}	$S_7 + S_9 \longrightarrow S_8$	$c_{17} \cdot X_7(t) \cdot X_9(t)$	k_{17}/Ω

Table 4.2: Reactions of the chemical reaction system. This table represents reactions of the chemical reaction system shown in Figure 4.1. The chemical reaction system includes $M = 17$ reactions. All reactions follow reaction types which are defined in Equation (4.2) and correspond to mass action kinetics in the deterministic model formulation [90]. Reactions R_1, R_5 and R_9 are inflow reactions. The majority of reactions is unimolecular, including conversion (R_3, R_7, R_{11} and R_{15}) and outflow reactions (R_2, R_4, R_6, R_8, R_{10}, R_{12} and R_{16}). Reactions R_{13}, R_{14} and R_{17} are bimolecular. Stochastic rates $c_1(t)$, $c_5(t)$ and $c_9(t)$ are time dependent and represent signal dependent mRNA productions. The stochastic rate constants (c_j in third column) and the deterministic rate constants (last column) are connected by $c_j = k_j\Omega$, $c_j = k_j$ and $c_j = k_j/\Omega$ with $\Omega = n_A \cdot vol$ and Avogadro's constant n_A for inflow, uni- and bimolecular reactions, respectively.

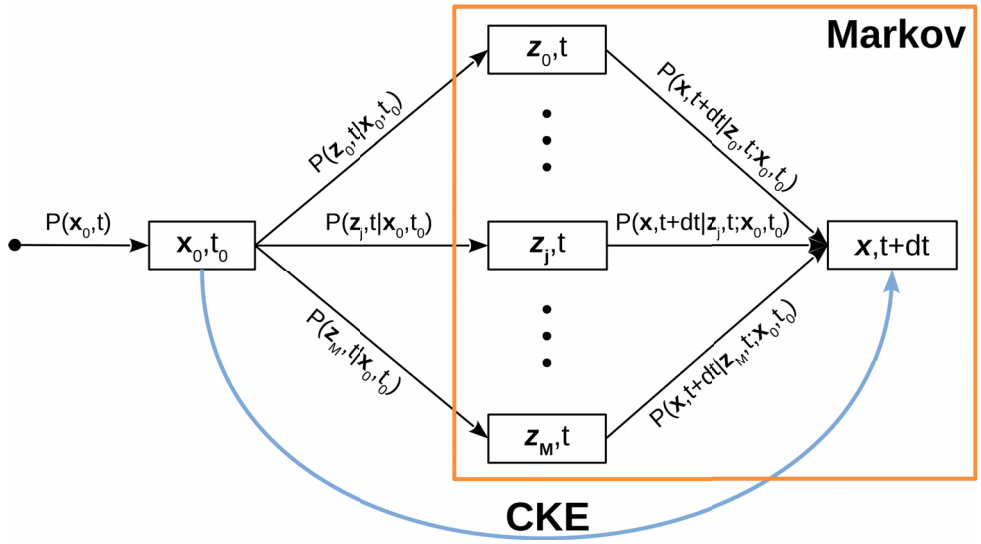

Figure 4.2: Reaction paths captured by the Chapman-Kolmogorov equation. In this figure, we show all possible $\boldsymbol{z}_j$ paths captured by the Chapman-Kolmogorov equation (CKE). The CKE describes the probability $P(\boldsymbol{x}, t+dt|\boldsymbol{x}_0, t_0)$ that the system is in final state $\boldsymbol{X}(t+dt) = \boldsymbol{x}$ given the system was in initial state $\boldsymbol{X}_0(t_0) = \boldsymbol{x}_0$ before (light blue arrow) by taking all possible intermediate states $\boldsymbol{Z}_j(t) = \boldsymbol{z}_j$ into account. The intermediate state $\boldsymbol{z}_0$ is equal to the final system state $\boldsymbol{x}$ and states $\boldsymbol{z}_j$ relates to M reactions which are defined in Table 4.2. Branch probabilities are conditional probabilities indicating the previous system state. The joint probability of e.g. the jth path is $P(\boldsymbol{x}, t+dt; \boldsymbol{z}_j, t; \boldsymbol{x}_0, t_0) = P(\boldsymbol{x}, t+dt|\boldsymbol{z}_j, t; \boldsymbol{x}_0, t_0)P(\boldsymbol{z}_j, t|\boldsymbol{x}_0, t_0)P(\boldsymbol{x}_0, t_0)$. The first term becomes $P(\boldsymbol{x}, t+dt|\boldsymbol{z}_j, t)$ if we look at a Markov process (orange box). Markov processes are characterized by regarding the most recent value of the process instead of evaluating the whole process history. The probabilities $P(\boldsymbol{x}, t+dt|\boldsymbol{z}_j, t)$ and $P(\boldsymbol{z}_j, t|\boldsymbol{x}_0, t_0)$ are named transition probabilities.

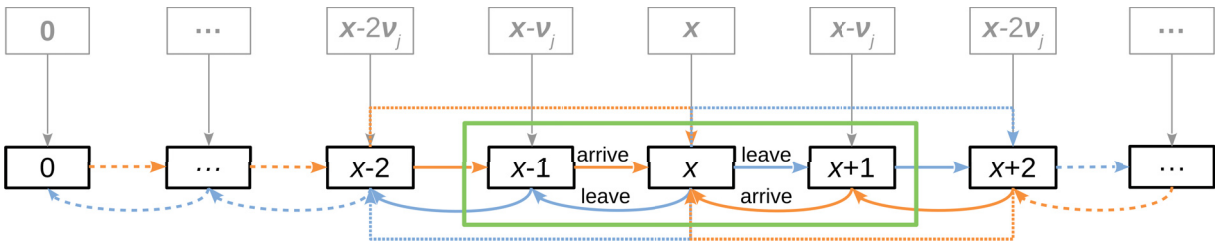

Figure 4.3: Processes covered by the chemical master equation. In this figure, we illustrate processes influencing $d_t P(\boldsymbol{x}, t|\boldsymbol{x}_0, t_0)$ which is the time evolution of the probability that there will be $\boldsymbol{X}(t) = \boldsymbol{x}$ molecules in the system volume V at time t, given the system was in state $\boldsymbol{X}_0(t_0) = \boldsymbol{x}_0$ before. We consider processes arriving at (solid orange arrows) or leaving (solid light blue arrows) $\boldsymbol{x}$ by just one reaction, named birth-death process (green boxe). To visualize all permitted states and processes, we show a simplified system $\varnothing \xrightarrow{c_1} S_1 \xrightarrow{c_2} \varnothing$ which includes inflow (R_1) and outflow (R_2) reactions of species S_1 (black boxes) and is characterized by state-change vectors $\nu_1 = +1$ and $\nu_2 = -1$. It is not relevant if c_1 is time dependent or not. If the system start from x, it can leave x by either producing $(x + \nu_1)$ or degrading $(x + \nu_2)$ one molecule. Similarly, if the system start from $x \pm 1$, it can arrive at x by either degrading $(x + \nu_2)$ or producing $(x + \nu_1)$ one molecule. Since we consider an infinitesimal time increment dt, where at most one reaction takes place, processes leading to or starting from states more than one reaction apart are not permitted (e.g. dotted arrows). In general, processes evaluating the most recent value only and ignoring the process history, no matter in which state the system actually is, are named Markov processes. Processes have a lower limit for $x = 0$ but no upper limit (dashed arrows). Since it is not possible to arrive at or leave $x = 0$ from or to a previous state (e.g. $x = -1$), $c_1 = c_2 = 0$ for these transitions. Gray boxes relates to the general case.

which can be transformed to

$$\boldsymbol{C}(t+\tilde{\tau}) = \boldsymbol{C}(t) + \tilde{\tau}\sum_{j=1}^{M}\boldsymbol{\nu}_j a_j(\boldsymbol{C}(t)) + \sqrt{\tilde{\tau}}\sum_{j=1}^{M}\boldsymbol{\nu}_j\sqrt{a_j(\boldsymbol{C}(t))}\mathcal{N}_j(0,1) \tag{4.16}$$

by using the linear combination $\mathcal{N}(\mu,\sigma^2) = \mu + \sigma\mathcal{N}(0,1)$ [88]. $\mathcal{N}(0,1)$ is a standardized normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$. In the limit $\tilde{\tau} \to 0$, Equation (4.16) becomes a continuous time process

$$d\boldsymbol{C}(t) = \sum_{j=1}^{M}\boldsymbol{\nu}_j a_j(\boldsymbol{C}(t))dt + \sum_{j=1}^{M}\boldsymbol{\nu}_j\sqrt{a_j(\boldsymbol{C}(t))}dW_j(t) \tag{4.17}$$

where $W_j(t)$ corresponds to statistically independent Brownian motion. That way, we moved from a jump-type (discrete) to a continuous Markov process [88]. Equation (4.17) is known as chemical Langevin equation (CLE), a special case of a stochastic differential equation (SDE).

In the thermodynamic limit, the number of molecules X_i of species S_i and the system volume V go to infinity but ratio X_i/V remains constant (see Figure 4.4). Thus, the stochastic part in Equation (4.17) becomes negligible and the CLE reduces to the RRE

$$d_t\boldsymbol{C}(t) = \sum_{j=1}^{M}\boldsymbol{\nu}_j a_j(\boldsymbol{C}(t)). \tag{4.18}$$

That means that we always assume to be in the thermodynamic limit if we use the deterministic formulation of a chemical reaction system (see Figure 4.5).

The RRE is typically expressed in species concentrations where stochastic rate constants c_j become deterministic rate constants k_j (see Table 4.3). Species concentrations are measured in units $[mol/l]$ corresponding to $n_i(t)/V$, the amount of substance n_i at time t per volume V. The amount of substance $n_i(t)$ is equivalent to $n_i(t) = X_i(t)/n_A$, the number of molecules X_i at time t divided by Avogadro's constant $n_A \approx 6.022 \cdot 10^{23}\ [1/mol]$. It follows for the system volume $V = vol \cdot dm^3$ (absolute measure $\cdot$ liter)

$$\begin{aligned}\frac{n_i(t)}{vol\cdot dm^3} &= \frac{X_i(t)}{n_A}\frac{1}{vol\cdot dm^3}\\ C_i(t) &= \frac{X_i(t)}{\Omega}\end{aligned} \tag{4.19}$$

with $C_i(t) = n_i(t)/vol$ and $\Omega = n_A \cdot vol$.

Stochastic rate constants c_j directly relates to deterministic rate constants k_j for conversion and outflow reactions (see Equation (4.2b) and (4.2c)). To keep units right, stochastic rate constants c_j are transformed to $k_j\Omega$ for inflow (see Equation (4.2a)) and k_j/Ω for bimolecular reactions (see Equation (4.2d)). The relation between propensities $a_j(\boldsymbol{X}(t))$ and $a_j(\boldsymbol{C}(t))$ is given by

$$c_j = k_j\Omega \qquad (k_j = c_j/\Omega), \tag{4.20a}$$
$$c_j X_h(t) = k_j C_h(t)\Omega \qquad (k_j = c_j), \tag{4.20b}$$
$$c_j X_h(t) = k_j C_h(t)\Omega \qquad (k_j = c_j), \tag{4.20c}$$
$$c_j X_g(t) X_h(t) = k_j C_g(t) C_h(t)\Omega \qquad (k_j = c_j\Omega), \tag{4.20d}$$

where $X_g(t) = C_g(t)\Omega$ and $X_h(t) = C_h(t)\Omega$.

We can show that the time evolution of the expected state $\langle \boldsymbol{X}(t)\rangle$ becomes exactly the RRE (see Equation (4.18)) for inflow, outflow and conversion reactions (see Equation (4.2a), (4.2b) and (4.2c)) by using the CME in Equation (4.13). The expected state is given by

$$\langle \boldsymbol{X}(t)\rangle \;=\; \sum_{\boldsymbol{x}} \boldsymbol{x} P(\boldsymbol{x}, t|\boldsymbol{x}_0, t_0) \tag{4.21}$$

$d_t\boldsymbol{C}(t) = \sum_{j=1}^{M} \boldsymbol{\nu}_j a_j(\boldsymbol{C}(t))$						
$d_tC_1(t)]$	$=$	$k_1(t)$	$-$	$k_2C_1(t)$		
$d_tC_2(t)$	$=$	$k_3C_1(t)$	$-$	$k_4C_2(t)$	$-$	$k_{13}C_2(t)C_7(t)$
$d_tC_3(t)$	$=$	$k_5(t)$	$-$	$k_6C_3(t)$		
$d_tC_4(0t)$	$=$	$k_7C_3(t)$	$-$	$k_8C_4(t)$		
$d_tC_5(t)$	$=$	$k_9(t)$	$-$	$k_{10}C_5(t)$		
$d_tC_6(t)$	$=$	$k_{11}C_5(t)$	$-$	$k_{12}C_6(t)$	$-$	$k_{13}C_2(t)C_7(t)$
$d_tC_7(t)$	$=$	$k_{13}C_2(t)C_7(t)$	$-$	$k_{14}C_4(t)C_7(t)$	$-$	$k_{17}C_7(t)C_9(t)$
$d_tC_8(t)$	$=$	$k_{14}C_4(t)C_7(t)$	$+$	$k_{17}C_7(t)C_9(t)$	$-$	$k_{15}C_8(t)$
$d_tC_9(t)$	$=$	$k_{15}C_8(t)$	$-$	$k_{16}C_9(t)$		

Table 4.3: RRE of the chemical reaction system. This table shows RRE for species concentrations $C_i(t)$ (see Table 4.1) according to reactions given in Table 4.2.

or

$$\langle f(\boldsymbol{X}(t))\rangle \quad = \quad \sum_{\boldsymbol{x}} f(\boldsymbol{x}) P(\boldsymbol{x}, t|\boldsymbol{x}_0, t_0) \tag{4.22}$$

where $\sum_{\boldsymbol{x}}$ is the sum over all $\boldsymbol{x} = (x_1, ..., x_N) = (X_1(t), ..., X_N(t))$ ranging element-wise from 0 to ∞. Therefore, the time derivative is

$$d_t\langle \boldsymbol{X}(t)\rangle = \sum_{j=1}^{M} \boldsymbol{\nu}_j \langle a_j(\boldsymbol{X}(t))\rangle \tag{4.23}$$

(see Appendix D.2 for calculation details). For linear functions holds

$$\langle f(\boldsymbol{X}(t))\rangle = f(\langle \boldsymbol{X}(t)\rangle) \tag{4.24}$$

wherefore Equation (4.23) becomes

$$d_t\langle \boldsymbol{X}(t)\rangle = \sum_{j=1}^{M} \boldsymbol{\nu}_j a_j(\langle \boldsymbol{X}(t)\rangle) \tag{4.25}$$

which is Equation (4.18) for the expected state $\langle \boldsymbol{X}(t)\rangle$. Since Equation (4.24) does not apply for bimolecular reactions, it is only the average of a large number $\mathcal{R}$ of numerical realizations $\boldsymbol{X}^{SSA}(t)$ of a large number of molecules which approximates the RRE in Equation (4.18) (see Figures 4.5 and E.1):

$$\overline{X_i^{SSA}(t)} = \frac{1}{\mathcal{R}} \sum_{r=1}^{\mathcal{R}} X_{ir}^{SSA}(t), \forall i. \tag{4.26}$$

A numerical realization $\boldsymbol{X}^{SSA}(t)$ is a random sample of $\boldsymbol{X}(t)$ and not the solution of the CME which is a probability density function $P(\boldsymbol{x}, t|\boldsymbol{x}_0, t_0)$.

4.4 Using Gillespie's stochastic simulation algorithm

The SSA, also known as Gillespie's algorithm, is an approach to calculate a numerical realization $\boldsymbol{X}^{SSA}(t)$ of the system state $\boldsymbol{X}(t)$ and is based on the same fundamental hypothesis as the CME (see Section 4.2). To derive the algorithm we introduce the reaction probability density function $p(\tau; j|\boldsymbol{x}, t)$ defined as

> $p(\tau; j|\boldsymbol{x}, t) \triangleq$ the probability that the next reaction in the system volume V will occur in the infinitesimal time interval $[t + \tau, t + \tau + d\tau)$ and will be reaction R_j given $\boldsymbol{X}(t) = \boldsymbol{x}$.

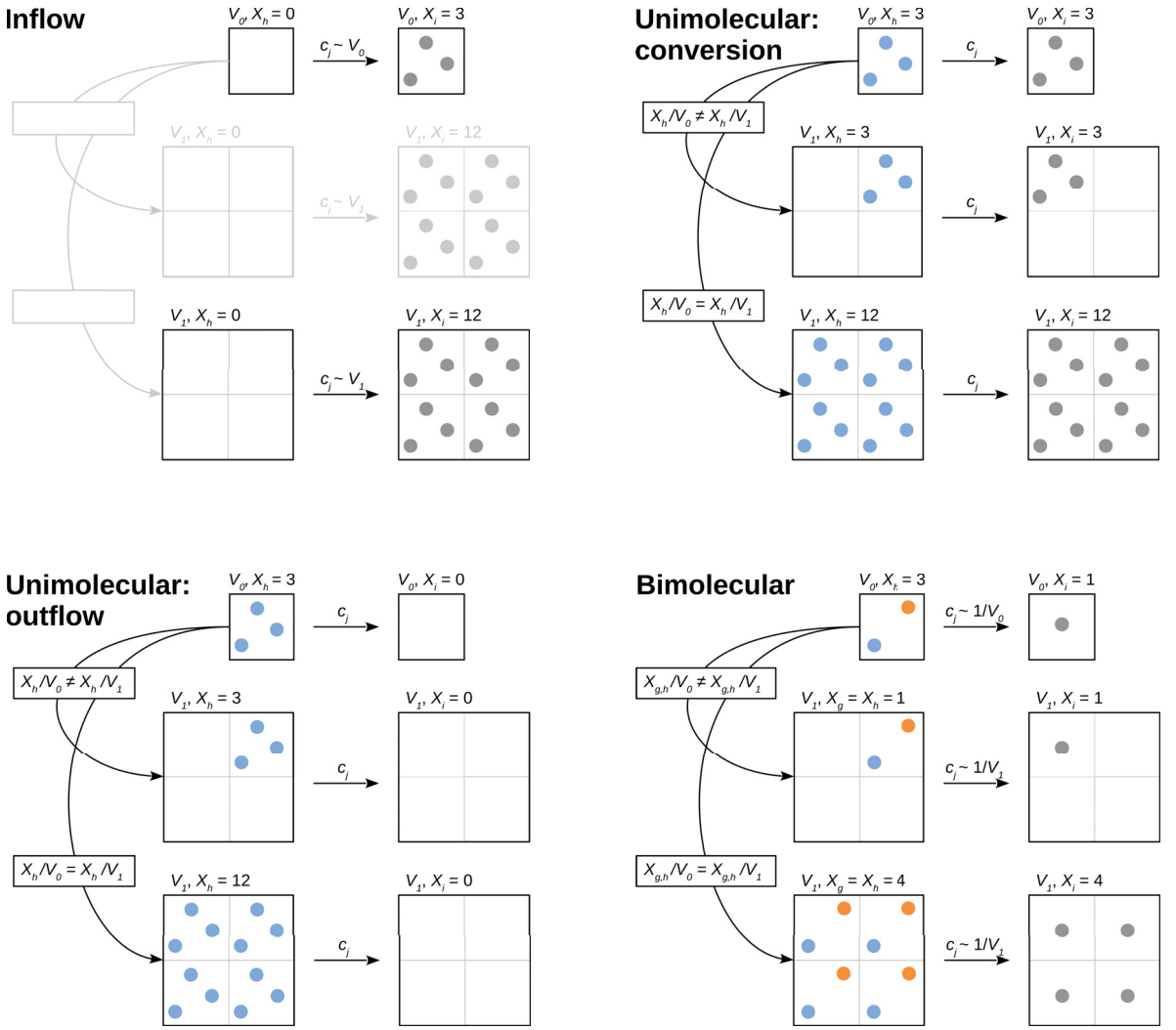

Figure 4.4: Thermodynamic limit. In this figure, the relation between the number of molecules X_i and the system volume V for inflow, uni- and bimolecular reactions (see Equations (4.2)) in the thermodynamic limit is illustrated. In addition, dependencies from the system volume V are shown for each reaction. The light blue, orange and gray molecules represent species S_h, S_g and S_i, respectively. If we increase the system volume V from V_0 to V_1 without increasing the number of molecules X_h, the ratio X_h/V_0 is not equal to the ratio X_h/V_1. If we increase the system volume V and the number of molecules X_h, the ratio X_h/V_0 is equal to the ratio X_h/V_1. In the thermodynamic limit where X_h and V go to infinity ($X_h \to \infty$, $V \to \infty$), the ratio X_h/V remains constant ($X_h/V = const.$). The same holds for the number of molecules X_g. Stochastic rate constants c_j are directly proportional to the system volume V for inflow reactions. The larger the system volume the more molecules are produced. These reactions are not influenced by the thermodynamic limit. Unimolecular reactions have stochastic rate constants c_j which are independent of the system volume V. The conversion from X_h molecules before to X_i molecules after or the outflow of X_h molecules can always take place. It is irrelevant where individual molecules stay in the system and how large the system is. A molecule only needs itself to convert or degrade. Stochastic rate constants c_j are inversely proportional to the system volume V for bimolecular reactions. The reaction from X_g and X_h molecules before to X_i molecules after can only take place if two molecules of each species collide. A collision is less likely in a larger system and it takes more time.

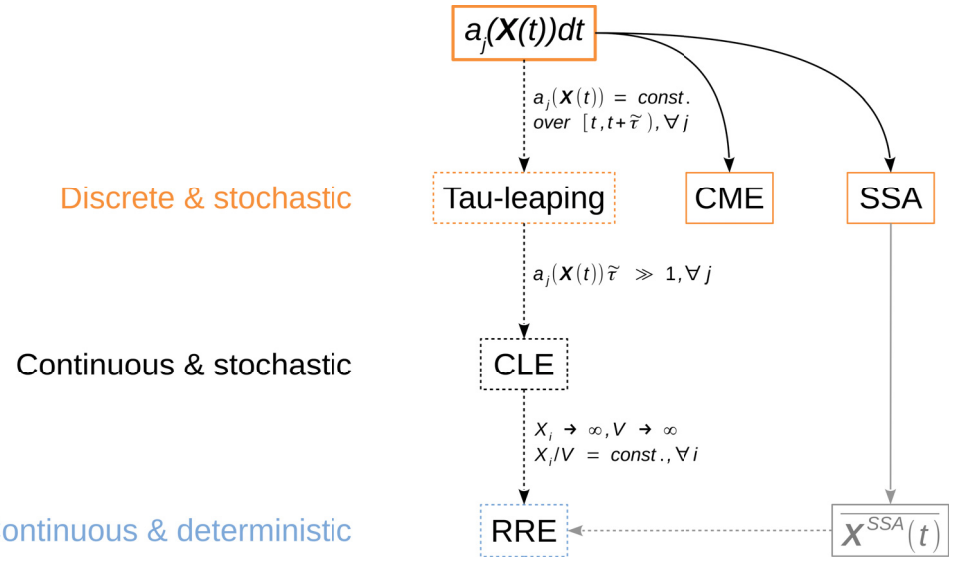

Figure 4.5: Relationship between the stochastic and deterministic formulation of a chemical reaction system. The fundamental hypothesis of the average probability $c_j dt$ that a particular combination of R_j reactant molecules will react in the next infinitesimal time interval $[t, t+dt)$ leads to the key quantity to describe a chemical reaction system. The propensity function $a_j(\boldsymbol{X}(t))$ is defined by the probability $a_j(\boldsymbol{X}(t))dt$ that one reaction R_j will occur in the next infinitesimal time interval $[t, t+dt)$ (solid bold orange box). The chemical master equation (CME) and the stochastic simulation algorithm (SSA) are directly derived from the fundamental hypothesis (solid orange boxes). Whereas the CME describes the time evolution of the probability $P(\boldsymbol{x}, t|\boldsymbol{x}_0, t_0)$, the SSA gives a numerical realization $\boldsymbol{X}^{SSA}(t)$ which represents a random sample $\boldsymbol{X}(t)$ of the stochastic process. The first step to derive the deterministic from the stochastic formulation is to introduce leap time $\tilde{\tau} > 0$. The leap time $\tilde{\tau}$ has to be so small that only a few reactions R_j take place in the time interval $[t, t+\tilde{\tau})$ and that propensities $a_j(\boldsymbol{X}(t))$ do not change significantly. Thus, the number of times reaction R_j fires in the time interval $[t, t+\tilde{\tau})$ becomes a pure counting process and a Poisson random variable (dotted orange box). The next assumption is $a_j(\boldsymbol{X}(t))\tilde{\tau} \gg 1$ so that each reaction R_j fires many more times than once. At this point, we switch to a continuous normal random variable and the chemical Langevin equation (CLE) is derived (dotted black box). The last step is to force the system into the thermodynamic limit ($X_i \to \infty$, $V \to \infty$ and $X_i/V = const.$) where the CLE reduces to the deterministic reaction rate equation (RRE) (dotted light blue box). As long as propensities are linear and it holds $\langle a_j(\boldsymbol{X}(t))\rangle = a_j(\langle \boldsymbol{X}(t)\rangle)$, the time evolution of the expected state is equivalent to the RRE. The relationship holds true even if we consider small molecule numbers. For non-linear propensity function where $\langle a_j(\boldsymbol{X}(t))\rangle \neq a_j(\langle \boldsymbol{X}(t)\rangle)$, the RRE can be approximated with the average of a large number $\mathcal{R}$ of realization $\boldsymbol{X}^{SSA}(t)$ of a large number of molecules where $\overline{X_i^{SSA}(t)} = \frac{1}{\mathcal{R}} \sum_{r=1}^{\mathcal{R}} X_{ir}^{SSA}(t), \forall i$ (solid light blue box). The average is an estimator for the expected state. Solid arrows indicate exact transitions whereas dotted arrows mark approximations. The figure structure is taken form [87] and adjusted.

The time increment $d\tau$ is so small that at most one reaction takes place. The reaction probability density function $p(\tau; j|\boldsymbol{x}, t)$ is formally a joint probability function for random variables τ (time until next reaction) and j (next reaction index) and arises from the joint probability function $p(\tau; j; \boldsymbol{x}, t)$ in the following way:

$$\begin{aligned} p(\tau; j; \boldsymbol{x}, t) = p(\boldsymbol{x}, t)p(\tau; j|\boldsymbol{x}, t) = p(\boldsymbol{x}, t)p(\tau|\boldsymbol{x}, t)p(j|\boldsymbol{x}, t; \tau) \\ p(\tau; j|\boldsymbol{x}, t) = p(\tau|\boldsymbol{x}, t)p(j|\boldsymbol{x}, t; \tau). \end{aligned} \tag{4.27}$$

At first, we assume that τ and j are independent events to express the reaction probability density function as

$$p(\tau; j|\boldsymbol{x}, t) = P_{01}(\tau|\boldsymbol{x}, t)P_{02}(j|\boldsymbol{x}, t) \tag{4.28}$$

the product of probability $P_{01}(\tau|\boldsymbol{x}, t)$ that no reaction takes place in the time interval $[t, t + \tau)$ and the subsequent probability $P_{02}(j|\boldsymbol{x}, t)$ that reaction R_j takes place in the time interval $[t + \tau, t + \tau + d\tau)$ with

$$P_{02}(j|\boldsymbol{x}, t) = a_j(\boldsymbol{x}). \tag{4.29}$$

To derive probability $P_{01}(\tau|\boldsymbol{x}, t)$, we make use of probability $P_{01}(\tau + d\tau|\boldsymbol{x}, t)$ that no reaction occurs in the time interval $[t, t + \tau + d\tau)$. The probability reads

$$P_{01}(\tau + d\tau|\boldsymbol{x}, t) = P_{01}(\tau|\boldsymbol{x}, t)\left(1 - \sum_{j=1}^{M} a_j(\boldsymbol{x})d\tau\right) \tag{4.30}$$

by assuming that probabilities $P_{01}(\tau|\boldsymbol{x}, t)$ and $\left(1 - \sum_{j=1}^{M} a_j(\boldsymbol{x})d\tau\right)$ are independent. The probability $\left(1 - \sum_{j=1}^{M} a_j(\boldsymbol{x})d\tau\right)$ that no reaction occurs in the time interval $[t+\tau, t+\tau+d\tau)$ is calculated from probability $\sum_{j=1}^{M} a_j(\boldsymbol{x})d\tau$ that any reaction fires in the time interval $[t + \tau, t + \tau + d\tau)$. Rearrangement leads to

$$\frac{P_{01}(\tau + d\tau|\boldsymbol{x}, t) - P_{01}(\tau)|\boldsymbol{x}, t)}{d\tau} = P_{01}(\tau|\boldsymbol{x}, t)\sum_{j=1}^{M} a_j(\boldsymbol{x}) \tag{4.31}$$

and results for $d\tau \to 0$ in

$$d_\tau P_{01}(\tau|\boldsymbol{x}, t) = P_{01}(\tau|\boldsymbol{x}, t)a_0(\boldsymbol{x}) \tag{4.32}$$

with $a_0(\boldsymbol{x}) = \sum_{j=1}^{M} a_j(\boldsymbol{x})$. This ODE has the initial condition $P_{01}(0|\boldsymbol{x}, t) = 1$ and the solution becomes

$$P_{01}(\tau|\boldsymbol{x}, t) = e^{-a_0(\boldsymbol{x})\tau}. \tag{4.33}$$

Inserting Equation (4.33) and (4.29) into Equation (4.28) gives

$$p(\tau; j|\boldsymbol{x}, t) = a_j(\boldsymbol{x})e^{-a_0(\boldsymbol{x})\tau}. \tag{4.34}$$

In a different ansatz, we assume that τ and j are dependent events to express the reaction probability density function as

$$p(\tau; j|\boldsymbol{x}, t) = P_1(\tau|\boldsymbol{x}, t)P_2(j|\boldsymbol{x}, t; \tau) \tag{4.35}$$

the product of probability $P_1(\tau|\boldsymbol{x}, t)$ that the next reaction will occur in the time interval $[t + \tau, t + \tau + d\tau)$ and probability $P_2(j|\boldsymbol{x}, t; \tau)$ that the next reaction at time $t + \tau$ will be reaction R_j. Probability $P_1(\tau|\boldsymbol{x}, t)$ is caluclated as the sum of $p(\tau, j|\boldsymbol{x}, t)$ over all M reactions

$$P_1(\tau|\boldsymbol{x}, t) = \sum_{j=1}^{M} p(\tau; j|\boldsymbol{x}, t) \tag{4.36}$$

so that probability $P_2(j|\boldsymbol{x}, t; \tau)$ follows from Equation (4.35)

$$P_2(j|\boldsymbol{x}, t; \tau) = \frac{p(\tau; j|\boldsymbol{x}, t)}{\sum_{j'=1}^{M} p(\tau; j'|\boldsymbol{x}, t)}. \tag{4.37}$$

Inserting Equation (4.34) into Equation (4.36) and (4.37) gives

$$P_1(\tau|\boldsymbol{x}, t) = a_0(\boldsymbol{x})e^{-a_0(\boldsymbol{x})\tau} \tag{4.38}$$

and

$$P_2(j|\boldsymbol{x}, t; \tau) = \frac{a_j(\boldsymbol{x})}{a_0(\boldsymbol{x})} \tag{4.39}$$

with $a_0(\boldsymbol{x}) = \sum_{j'=1}^{M} a_{j'}(\boldsymbol{x})$. Both probability density functions are normalized

$$\int_0^\infty P_1(\tau|\boldsymbol{x}, t)d\tau = \int_0^\infty a_0(\boldsymbol{x})e^{-a_0(\boldsymbol{x})\tau}d\tau = 1 \tag{4.40}$$

$$\sum_{j=1}^{M} P_2(j|\boldsymbol{x}, t; \tau) = \sum_{j=1}^{M} \frac{a_j(\boldsymbol{x})}{a_0(\boldsymbol{x})} = 1. \tag{4.41}$$

In Equation (4.38) we see that τ is an exponential random variable with mean and standard deviation $1/a_0(\boldsymbol{x})$. Contrarily, Equation (4.39) shows that $P_2(j|\boldsymbol{x}, t; \tau)$ only depends on j which is an integer random variable with point probability $a_j(\boldsymbol{x})/a_0(\boldsymbol{x})$.

Finally, the joint probability density of two random variables, the time until next reaction τ (continuous variable: $0 \leq \tau < \infty$) and the next reaction index j (discrete variable: $j = 1, ..., M$), split into two probability density functions of just one random variable and reads

$$\begin{aligned} p(\tau, j|\boldsymbol{x}, t) &= P_1(\tau|\boldsymbol{x}, t)P_2(j|\boldsymbol{x}, t; \tau) \\ &= \frac{a_j(\boldsymbol{x})}{a_0(\boldsymbol{x})} a_0(\boldsymbol{x})e^{-a_0(\boldsymbol{x})\tau}. \end{aligned} \tag{4.42}$$

In the SSA, we make use of this property of the reaction probability density function by sampling τ and j separately to get a random sample pair (τ, j). Monte Carlo sampling is used to generate a random sample of each random variable. In the direct method, two random numbers ξ_1 and ξ_2 are sampled independently from the uniform distribution in the unit interval $\mathcal{U}(0, 1)$ and subsequently transformed into a random time τ or a random index j by inversion. Thus, we can sample all possible values of τ and j with the same probability even if they are not uniformly distributed (see Figure 4.6).

The following steps illustrate how a numerical realization $\boldsymbol{X}^{SSA}(t)$ of the stochastic process is generated:

1. System initialization: $t = t_0$ and $\boldsymbol{x} = \boldsymbol{x}_0$.
2. Propensity evaluation: $\{a_j(\boldsymbol{x})\}_{j=1}^{M}$ and $a_0(\boldsymbol{x})$.
3. $\mathcal{U}(0, 1)$ sampling: ξ_1 and ξ_2.
4. ξ_1 inversion: $\tau = \frac{1}{a_0(\boldsymbol{x})}\ln(\frac{1}{\xi_1})$.
5. ξ_2 inversion: $j =$ the smallest integer satisfying $\xi_2 a_0(\boldsymbol{x}) \leq \sum_{j'=1}^{j} a_{j'}(\boldsymbol{x})$.
6. System update: $t = t + \tau$ and $\boldsymbol{x} = \boldsymbol{x} + \boldsymbol{\nu}_j$.
7. $\boldsymbol{X}^{SSA}(t)$ recording.
8. Final decision: return to step 2 or stop simulation.

Practically, it is not possible to compare a number of realizations with different event times with real data (e.g. distribution at a specific time point) or calculate the average over time. The reason is that you never know in advance how many time steps will occur in a fixed time interval. Therefore, we used a discretized version of the SSA with a regular time grid of predefined time steps. We used the R function `gillespied()` of the R package **smfsb**[7]. A further advantage of the discretized SSA is that we do not have to deal with inversions of time dependent stochastic rates defined for mRNA productions in step 4 (see Table 4.2 and Appendix D.3 for calculation details).

[7] freely available on `https://CRAN.R-project.org/package=smfsb`

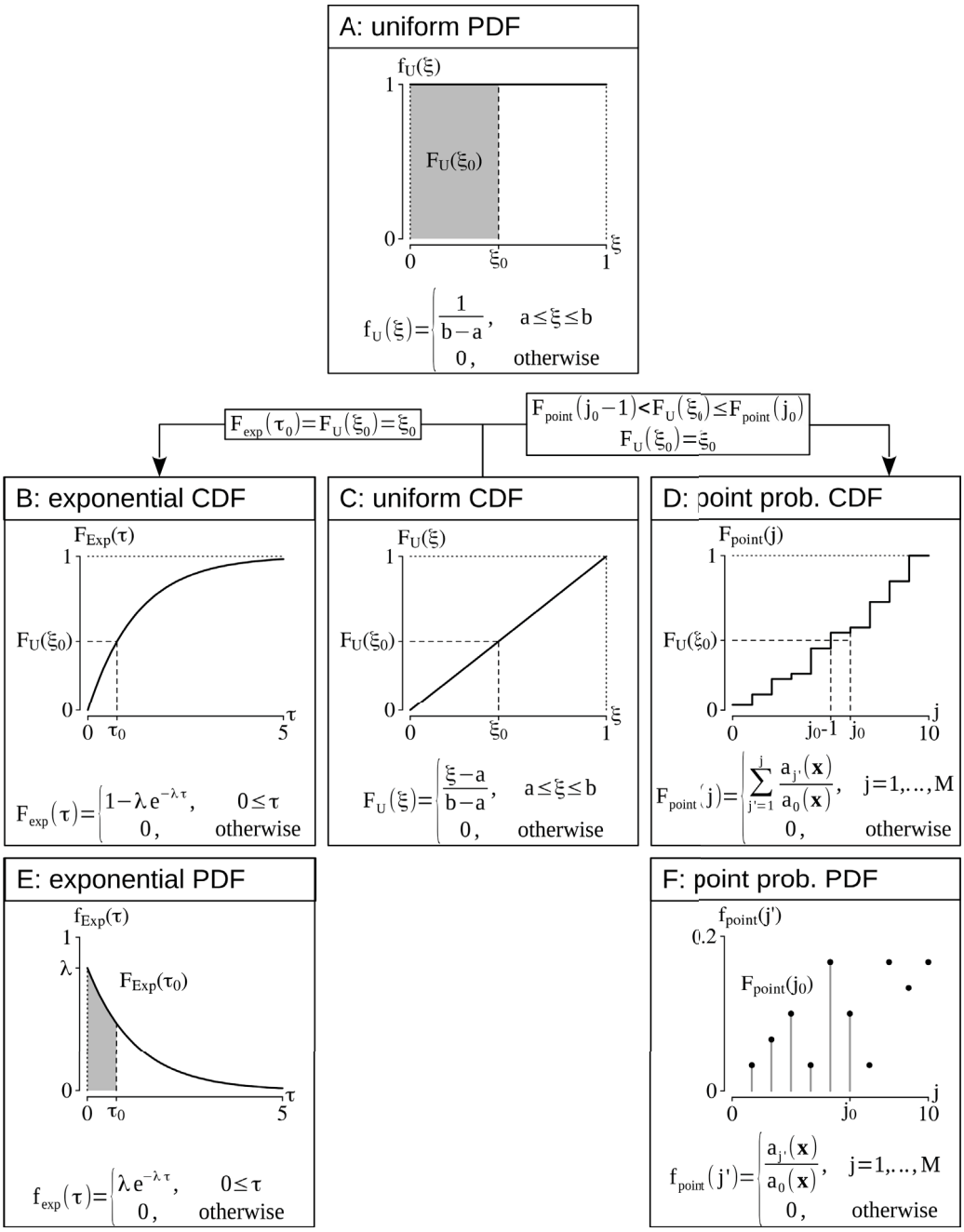

Figure 4.6: Monte Carlo sampling by standard inversion. The aim of Monte Carlo sampling is to draw a random number from a not uniformly distributed random variable with equal probability for each possible random number. In our case, the time until next reaction τ is exponentially distributed ($f_{Exp}(\tau)$ in E with $\lambda = a_0(\boldsymbol{x})$) and the next reaction index j follows a discrete probability density function (PDF) representing point probabilities ($f_{point}(j')$ in F). First, we draw a random number ξ_0 from the uniform distribution in the unit interval ($f_U(\xi)$ in A with $a = 0$ and $b = 1$). Each possible value for ξ_0 is equally likely. The probability of ξ to be at most ξ_0 in the unit interval is given by $F_U(\xi_0) = \xi_0$ (gray area in A and dashed line in C) where $F_U(\xi) = \int_{-\infty}^{\xi} f_U(\xi)d\xi$ is the cumulative distribution function (CDF). Since our probability density functions are normalized ($F_U(\xi) = \int_{-\infty}^{\infty} f_U(\xi)d\xi = 1$, $\int_{-\infty}^{\infty} f_{Exp}(\tau)d\tau = 1$ and $\sum_{j'=1}^{M} f_{point}(j') = 1$), we can set $F_U(\xi_0) = F_{Exp}(\tau_0) = \xi_0$ with τ_0 the time of interest (left branch; gray area in A is equal to gray area in E) as well as require $F_{point}(j_0 - 1) < F_U(\xi_0) \le F_{point}(j_0)$ with $F_U(\xi_0) = \xi_0$ and j_0 the reaction index of interest (right branch; gray area in A is equal to gray bars in F). Here, $F_{Exp}(\tau)$ and $F_{point}(j)$ are the cumulative distribution function $F_{Exp}(\tau) = \int_{-\infty}^{\tau} f_{Exp}(\tau)d\tau$ and $F_{point}(j) = \sum_{j'}^{j} f_{point}(j')dj'$. Now, we invert our cumulative distribution functions to get τ_0 and j_0 out of $F_U(\xi_0) = \xi_0$. The time until next reaction becomes $\tau_0 = (1/\lambda)\ln(1/(1-\xi_0))$ with $\lambda = a_0(\boldsymbol{x})$ which is statistically equivalent to $\tau_0 = (1/\lambda)\ln(1/\xi_0)$. The next reaction index j_0 has to satisfy $\sum_{j'=1}^{j_0-1} a_{j'}(\boldsymbol{x}) < a_0(\boldsymbol{x})\xi_0 \le \sum_{j'=1}^{j_0} a_{j'}(\boldsymbol{x})$ which is statistically equivalent to the smallest value fulfilling $a_0(\boldsymbol{x})\xi_0 \le \sum_{j'=1}^{j_0} a_{j'}(\boldsymbol{x})$.

4.5 Representation of the experimental data by the stochastic model

We use a simplified system which describes the production and the degradation of species S_1, $\varnothing \xrightarrow{c_1} S_1 \xrightarrow{c_2} \varnothing$, to show how experimental data are represented by the stochastic model. The stationary distribution $P^\star(x,t)$ of the birth-death process is a Poisson distribution $\mathcal{P}(\lambda^\star)$ with $\lambda^\star = c_1/c_2$ (see Appendix D.4 for calculation details and Figure E.2). It is not relevant if the production rate is a stochastic rate constant c_1 or a time dependent stochastic rate $c_1(t)$ as long as the limit $\lim_{t\to\infty} c_1(t)$ exists.

smFISH data

smFISH data reveal mRNA distributions per cell cycle phase (see Section 3.1). Cell cycle phases are specific time intervals. The stochastic process delivers distributions per time point. Therefore, we consider the average distribution over all time points per time interval. As long as the production rate is a stochastic rate constant c_1, the average distribution of a time interval (TI) in the steady state region is exactly the same Poisson distribution as at any time point (TP) within the time interval:

$$\mathcal{P}_{TP}(\lambda^\star) = \mathcal{P}_{TI}(\lambda^\star) = \frac{1}{n_t} \sum_{t \in TI} \mathcal{P}_{TP}(\lambda^\star) \tag{4.43}$$

where n_t is the number of time points within the time interval (see Figure E.2). The Poisson distribution in a time interval is better described because we consider more molecule counts. The average distribution of a time interval outside the steady state region is not well approximated by a Poisson distribution even if the distribution at any time point within the time interval is Poisson distributed:

$$\mathcal{P}_{TP}(\lambda(t)) \neq \frac{1}{n_t} \sum_{t \in TI} \mathcal{P}_{TP}(\lambda(t)) \tag{4.44}$$

where $\lambda(t)$ describes the varying mean value of the Poisson distribution over time. The fact that the distribution outside the steady state region at a specific time point is still Poisson distributed comes from the birth-death process itself [95].

The behavior of a time dependent stochastic rate $c_1(t)$ depends on the underlying function. If $c_1(t)$ has no limit, e.g. a linear function, the system never reaches a steady state and, therefore, has no stationary distribution. Apart from very small time intervals, the average distribution will always be different from the Poisson distribution at any time point within the time interval. If $c_1(t)$ has a limit, e.g. a sigmoid function, the behavior is comparable to the case of a stochastic rate constant.

We use a time dependent stochastic rate which is represented by a step function. At this point it is not relevant if transitions between high/low and low/high stochastic rate constants is continuous or discontinuous as defined in Table 4.2. We consider a single cell cycle passage. Nevertheless, the change between high and low stochastic rate constants is recurring over several passages why the system never reaches a steady state.

In low transcription regions, the average distribution of a time interval is the same Poisson distribution as at any time point within the time interval. The system behave like being in steady state (see Figure 4.7). The same situation can arise in high transcription regions if a maximum level persists for a certain time. In regions of increasing/decreasing molecule numbers, the system behave like a typical non steady state system. The average distribution of a time interval is not the same Poisson distribution as at any time point within the time interval.

The most critical case is a time interval around the transition point (see Figure 4.7). Here, the average distribution is completely different to any Poisson distribution inside the time interval. The average distribution is a sum of Poisson distributions with varying mean values $\lambda_{low}(t)$ and

$\lambda_{high}(t)$:

$$\mathcal{P}_{TP}(\lambda(t)) \neq \frac{1}{n_t}\left(\sum_{t\in[0,t_{s_{1,0}}]\vee[t_{s_{1,e}},0]} \mathcal{P}_{TP}(\lambda_{low}(t)) + \sum_{t\in[t_{s_{1,0}},t_{s_{1,e}},0]} \mathcal{P}_{TP}(\lambda_{high}(t))\right). \tag{4.45}$$

The influence of low and high mean valued Poisson distributions on the shape of the final distribution depends on the position of the time interval and, therefore, on the weighting of these Poisson distributions which is similar to a two component Poisson distribution. There can be either two clear maxima or only one.

In fact, the average distribution over a time interval for a step function is only in some cases a Poisson distribution and, hence, suited to a limited extent to compare model output with our measured mRNA distributions which we assume to be Poisson distributed. Special attention is required if the average distribution is calculated around the transition states. Nevertheless, the average distribution is a reasonable approximation for distributions per time interval.

Western blot data

Since we do not have any measured distributions for the number of proteins, we assume that measured protein time courses represent the time evolution of the system's expected state. In Section 4.3 we derived the RRE (see Equation (4.18)) from the CME (see Equation 4.13) which is exactly the time evolution of the expected state (see Equation (4.25)) for inflow and unimolecular reactions (see Figure 4.5). This is the case for most reactions of the protein part.

The RRE can also be approximated by the average of a large number of numerical realizations of a large number of molecules (see Equation (4.26)) for bimolecular reactions (see Figure E.1). Given by the measurement technique, it is reasonable to assume that protein numbers are large, so that the approximation becomes better with the number of realizations. Thus, we use the RRE to compare the stochastic model output with measured protein time courses.

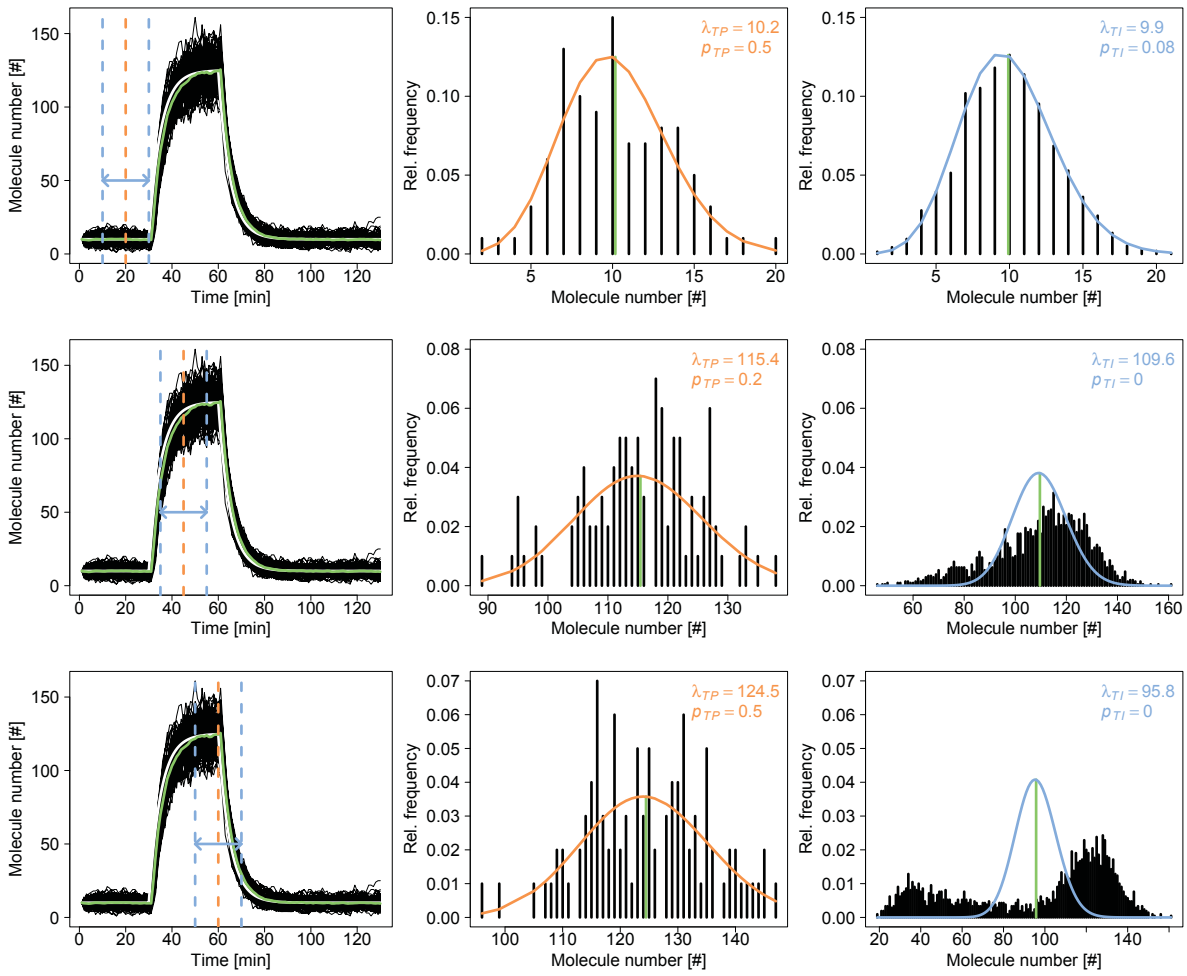

Figure 4.7: Time point and time interval distributions for a time dependent mRNA production rate defind by a step function. In this figure, 100 realizations (solid black lines) of the production and the degradation of species S_1, $\varnothing \xrightarrow{c_1(t)} S_1 \xrightarrow{c_2} \varnothing$, are shown to compare the average distribution of a time interval (TI, light blue) with the distribution at any time point within the time interval (TP, orange). The production rate is a time dependent stochastic rate $c_1(t)$ as defined in Table 4.2. In a recurring change between high and low transcription regions in several successive passages through the cell cycle, the system will never reaches a steady state. The first row shows the same average distribution of a time interval as the Poisson distribution at any time point within the time interval. The system behaves like being in steady state. The estimated mean values, λ_{TP} and λ_{TI}, show small differences resulting from the simulation itself. In the region of increasing molecule numbers (middle row), both distributions cleary differ and the average distribution is no longer well described by a Poisson distribution. In the last row, the time interval is positioned around the transition state. The time point distribution is Poisson distributed but the average distribution is influenced by time point Poisson distributions from the high and the low transcription region. We used a χ^2 goodness of fit test to calculate p-values (see Appendix D.9 for calculation details). Large p-values compared to a significance level of $\alpha = 0.05$ indicate distributions reasonably represented by a Poisson distribution. Lines between relative frequencies of Poisson distributions are for visualization only. All given values are rounded to the first decimal different from zero. The mean value of all realizations $\overline{X^{SSA}(t)}$ is represented by a light green solid line and the solution of the corresponding RRE by a white solid line.

5. Parameter estimation

All definitions in this chapter are taken from [61, 63, 77, 96, 97, 98]. Different references are indicated.

5.1 Estimating parameters in ODE systems

We use parameter estimation methods developed for non-linear dynamical systems which are represented by ODEs and based on time series or dose response data. Starting from Equation (4.18), we now consider an explicit dependency of the system's variables $\boldsymbol{C}(t,\theta)$ and the propensity function $\boldsymbol{a}(\boldsymbol{C}(t,\theta),\theta)$ on the parameter vector $\theta = (k_1, ..., k_M, C_{1,0}, ..., C_{N,0})$. The parameter vector includes deterministic rate constants k_j and initial values $C_{i,0}$ of the initial state vector $\boldsymbol{C}(0,\theta) = \boldsymbol{C}_0 = (C_{1,0}, ..., C_{N,0})$. The equation system becomes

$$\begin{aligned} d_t\boldsymbol{C}(t,\theta) &= \boldsymbol{f}(\boldsymbol{C}(t,\theta),\theta) \\ &= \boldsymbol{\nu}\boldsymbol{a}(\boldsymbol{C}(t,\theta),\theta) \end{aligned} \tag{5.1}$$

where $\boldsymbol{f}$ is a function describing the temporal evolution of variables $\boldsymbol{C}(t,\theta)$, $\boldsymbol{\nu}$ is the stoichiometric matrix (see Table F.1) and $\boldsymbol{a}(\boldsymbol{C}(t,\theta),\theta)$ is the propensity vector (see Table 4.2).

In practice, not all variables are measurable and, therefore, the chemical reaction system is partially observable. We map variables $\boldsymbol{C}(t,\theta)$ to observables $\boldsymbol{Y}(t,\theta) = (Y_1(t,\theta), ..., Y_{nObs}(t,\theta))$ by using the observation function $\boldsymbol{g}$

$$\boldsymbol{Y}(t,\theta) = \boldsymbol{g}(\boldsymbol{C}(t,\theta),\theta). \tag{5.2}$$

Observables of our chemical reaction system are given in Table 5.1. The number of observables $nObs$ is smaller than the number of species N in the system. Experimental data $\boldsymbol{y} = (y_1, ..., y_{nObs})$ represent corresponding measurements for each observable

$$\boldsymbol{y} = \boldsymbol{Y}(t,\theta) + \boldsymbol{\varepsilon} \tag{5.3}$$

where ε is the measurement noise.

Parameter estimation is the process of inferring parameter values of a statistical model based on experimental data [63]. The maximum likelihood estimate (MLE)

$$\hat{\theta} = arg\ \max_{\theta}\ L(y_k|\theta) \tag{5.4}$$

is an estimator that makes use of the distribution of the measurement noise and provide a point estimate for parameters of interest.

The likelihood function $L(y_k|\theta)$ measures how likely the observed data y_k are, given a parameter vector θ

$$L(y_k|\theta) = \phi(y_k|\theta). \tag{5.5}$$

If all observations $y_k, ...y_{nObs}$ are iid samples of the same density $\phi(y_k|\theta)$, the likelihood of the combined data is

$$L(\boldsymbol{y}|\theta) = \prod_{k=1}^{nObs} \phi(y_k|\theta) \tag{5.6}$$

and the negative log-likelihood becomes

$$\begin{aligned} \ell(\boldsymbol{y}|\theta) &= -\log(L(\boldsymbol{y}|\theta)) \\ &= -\sum_{k=1}^{nObs} \log(\phi(y_k|\theta)). \end{aligned} \tag{5.7}$$

$\boldsymbol{Y}(t,\theta) = \boldsymbol{g}(\boldsymbol{C}(t,\theta),\theta)$								
$Y_1(t,\theta)$	=	$C_1(t,\theta)$						
$Y_2(t,\theta)$	=	$C_2(t,\theta)$	+	$C_7(t,\theta)$	+	$C_8(t,\theta)$		
$Y_3(t,\theta)$	=	$C_3(t,\theta)$						
$Y_4(t,\theta)$	=	$C_4(t,\theta)$						
$Y_5(t,\theta)$	=	$C_5(t,\theta)$						
$Y_6(t,\theta)$	=	$C_5(t,\theta)$	+	$C_7(t,\theta)$	+	$C_8(t,\theta)$	+	$C_9(t,\theta)$

Table 5.1: Observables of the chemical reaction system. This table shows the observables according to the chemical reaction system which is given in Table 4.2 and available data ($nObs = 6$). Observables $Y_k(t,\theta)$ for $k = \{1,3,4,5\}$ are equal to variables $C_i(t,\theta)$ whereas observables $Y_k(t,\theta)$ for $k = \{2,6\}$ are superpositions of variables $C_i(t,\theta)$.

The log-likelihood function is more convenient compared to the likelihood function itself due to analytical and computational reasons. Similarly, it is more common to minimize the negative log-likelihood function than maximizing the log-likelihood function. Both optimizations are equivalent because logarithmic functions are strictly monotonically increasing. Finally, the maximum likelihood estimate in Equation 5.4 turns into

$$\hat{\theta} = arg\ \min_{\theta}\ \ell(\boldsymbol{y}|\theta). \tag{5.8}$$

The negative log-likelihood function is what we call objective function. The optimization problem is non-linear and non-convex why the objective function has multiple local optima and global optimization methods are needed [99].

We used a 2-step-optimization to parameterize our chemical reaction system. First, mRNA parameters are estimated from smFISH data in the mRNA optimization step. Secondly, mRNA parameters are re-estimated and protein parameters are estimated from Western blot data in the protein optimization step. In both steps, we decided for a multi-start local (deterministic) optimization method. These methods are based on the gradient of the objective function $d_\theta \ell(\boldsymbol{y}|\theta)$ which requires the calculation of model sensitivities $d_\theta \boldsymbol{C}(t,\theta)$. The whole algorithm is implemented in the programming language R.

5.2 Applying the maximum likelihood approach

mRNA optimization step

Each mRNA species (S_i, $i = \{1,3,5\}$) evolve statistically independent and separated from the remaining chemical reaction system (see Tables 4.1 and 4.2). Consequently, we can estimate mRNA parameters separately for each mRNA species by using respective negative log-likelihood functions $\ell(y_k|\theta)$ and do not have to consider the negative log-likelihood function of the combined data $\ell(\boldsymbol{y}|\theta)$ (see Equation 5.7).

In addition, all gene transcripts follow the same reaction scheme

$$\varnothing \xrightarrow{c_1(t)} S_1 \xrightarrow{c_2} \varnothing \tag{5.9}$$

with an inflow and an outflow reaction as described in Equations 4.2a and 4.2b, here exemplified for species S_1. The analytical solution of the CME

$$\begin{aligned} d_t P(x,t|x_0,t_0) = {} & c_1(t)P(x-1,t|x_0,t_0) + c_2(x+1)P(x+1,t|x_0,t_0) \\ & - c_1(t)P(x,t|x_0,t_0) - c_2 x P(x,t|x_0,t_0) \end{aligned} \tag{5.10}$$

for a Poisson initial distribution

$$P(x,0) = \mathcal{P}(x,\lambda_0) \tag{5.11}$$

is still a Poisson distribution

$$\begin{aligned} P(x,t|x_0,t_0) &= \mathcal{P}(x,\lambda(t)) \\ &= \frac{\lambda(t)^x}{x!}e^{-\lambda(t)} \end{aligned} \tag{5.12}$$

which is parameterized by

$$d_t\lambda(t) = k_1(t) - k_2\lambda(t). \tag{5.13}$$

This ODE is equivalent to the RRE of the reduced model system (see Equation 5.1 and Appendix D.5 for calculation details) [95]. Since we assume that measured mRNA distributions are Poisson distributed, we can use Equation 5.13 to calculate observables (see Equation 5.2 and Table 5.1). For the analytical solution it is not relevant if the stochastic rate $c_1(t)$ is time dependent or a constant.

Now, we can simulate Equation 5.13 and calculate Poisson distributions at any time point $t \in [0, T_{cycle}]$. We can use these Poisson distributions together with the measured distributions to set a likelihood function. Experimental data y_k ($k = \{1, 3, 5\}$) are metrices according to the number of measured mRNA distributions per mRNA species.

The likelihood function for $y_{klm} \sim \mathcal{P}(\gamma_{klm})$ becomes

$$L_{mRNA}(y_k|\theta^k) = \prod_{l=1}^{nP}\prod_{m=1}^{nB}\frac{(\gamma_{klm})^{y_{klm}}}{y_{klm}!}e^{-\gamma_{klm}} \tag{5.14}$$

and the corresponding negative log-likelihood function is

$$\ell_{mRNA}(y_k|\theta^k) = \sum_{l=1}^{nP}\sum_{m=1}^{nB}\left[\gamma_{klm} + \log(y_{klm}!) - y_{klm}\log(\gamma_{klm})\right] \tag{5.15}$$

with

$$\gamma_{kl} = \frac{\mathcal{Z}_{kl}}{nT}\cdot\boldsymbol{\mathcal{P}}\cdot\mathbf{1} \tag{5.16}$$

and the specific parameter vector $\theta^k \subset \theta$ for observable Y_k [52]. nP is the number of cell cycle phases and nB the number of bins per distribution. Thus, y_{klm} and γ_{klm} are measured and simulated mRNA frequencies in cell cycle phase l and bin m. The matrix $\boldsymbol{\mathcal{P}}$ includes all simulated Poisson distributions and is defined as

$$\boldsymbol{\mathcal{P}} = \{\mathcal{P}_{an}\} = \left\{\frac{Y_k(t_{ln},\theta^k)^{A_{ka}}}{A_{ka}!}e^{-Y_k(t_{ln},\theta^k)}\right\} \tag{5.17}$$

where A_{ka} are measured numbers of mRNA molecules running from $a = 1, ..., nR$ and t_{ln} are time points in cell cycle phase l running from $n = 1, ..., nT$ (see Definitions C.3 and C.1 in Appendix C). The number of time points nT is dependent on the finite time step Δt which is used for numerical calculations of Equation 5.13. In practice, we first solve Equation 5.13 for all time points $t \in [0, T_{cycle}]$ before we determine $Y_k(t_{ln}, \theta^k)$ to avoid discontinuities at the joins of subintervals. Multiplying matrix $\boldsymbol{\mathcal{P}}$ by vector $\mathbf{1} = (1, 1, ...)$ of length nR results in a summation over all time points t_{ln} within a cell cycle phase l for all mRNA numbers A_{ka}. The normalization factor $\mathcal{Z}_{kl}$ is the number of measured cells (see Definition C.2 in Appendix C) and nT the number of time points in cell cycle phase l.

Optimization was done with an optimization method classified as line search algorithm which is implemented in the R function `optim()`. The R function uses a quasi-Newton method with box constraints ranging from zero to infinity which is named L-BFGS-B method [100, 101, 102, 103]. This constrained optimization ensures positively valued parameter estimates. Here, we calculated model sensitivities by using the finite difference approximation

$$d_{\theta_i^k}C_i(t_{ln},\theta^k) = \frac{C_i(t_{ln},\theta^k) - C_i(t_{ln},\theta^k + h\boldsymbol{e}_\iota)}{h} \tag{5.18}$$

where $\boldsymbol{e}_\iota$ is the ι-th unit vector and h the step control which has to be chosen sufficiently small (see Appendix D.6 for calculation details) [100]. We chose $h = 0.001$ in all calculations. Contrary to the definition in Table 4.2, we implemented continuous transitions between low $k_{j,low}$ and high $k_{j,high}$ mRNA production rates by using Fermi-Dirac distributions (see Appendix D.8). Further, we fixed initial value $\lambda_0 = 0$ because we do not expect any mRNA molecule at $t = 0$.

Protein optimization step

Protein species (S_i, $i \neq \{1, 3, 5\}$) do not evolve statistically independent. They depend on mRNA and protein species (see Tables 4.1 and 4.2). Thus, we have to estimate protein parameters in a combined optimization and consider the negative log-likelihood function of the combined data $\ell(\boldsymbol{y}|\theta)$ (see Equation 5.7). Further, parameters are log-transformed and exposed to a Gaussian prior to guarantee positively non-infinite parameter values [104]. Here, we measured protein time courses for each observable Y_k why experimental data y_k ($k = \{2, 4, 6\}$) are vectors.

The combined likelihood function for $y_{kn} = Y_k(t_n, \theta) + \varepsilon_{kn}$ with $\varepsilon_{kn} \sim \mathcal{N}(0, \sigma_{kn}^2)$ becomes

$$L_{protein}(\boldsymbol{y}|\theta) = \prod_{k=\{2,4,6\}} \prod_{n=1}^{nT} \frac{1}{\sqrt{2\pi\sigma_{kn}^2}} e^{-\frac{(y_{kn} - Y_k(t_n,\theta))^2}{2\sigma_{kn}^2}} \tag{5.19}$$

and the corresponding negative log-likelihood function is

$$\ell_{protein}(\boldsymbol{y}|\theta) = \sum_{k=\{2,4,6\}} \sum_{n=1}^{nT} \frac{1}{2} \left[\left(\frac{y_{kn} - Y_k(t_n, \theta)}{\sigma_{kn}} \right)^2 + \log(2\pi\sigma_{kn}^2) \right] \tag{5.20}$$

with nT the number of time points $t \in [0, T_{cycle}]$.

Prior knowledge can be added to the likelihood function by multiplying with the prior probability density function of the respective parameters giving a posterior probability density function [104, 105, 106, 107]. Treating the prior probability density function as "prior likelihood function", the posterior probability density function is another combined likelihood function

$$L'_{protein}(\boldsymbol{y}|\theta) = L_{protein}(\boldsymbol{y}|\theta) \cdot p_{Gauss}(\theta) \tag{5.21}$$

as seen in Equation 5.6 and the corresponding negative log-transform is

$$\ell'_{protein}(\boldsymbol{y}|\theta) = \ell_{protein}(\boldsymbol{y}|\theta) + \sum_{\iota=1}^{nPar} \frac{1}{2} \left[\left(\frac{\theta_\iota^* - \theta_\iota}{\sigma_{\theta_\iota}} \right)^2 + \log(2\pi\sigma_{\theta_\iota}^2) \right] \tag{5.22}$$

with

$$p_{Gauss}(\theta) = \prod_{\iota=1}^{nPar} \frac{1}{\sqrt{2\pi\sigma_{\theta_\iota}^2}} e^{-\frac{(\theta_\iota^* - \theta_\iota)^2}{2\sigma_{\theta_\iota}^2}} \tag{5.23}$$

for $\theta_\iota \sim \mathcal{N}(\theta_\iota^*, \sigma_{\theta_\iota}^2)$ where $nPar$ is the number of parameters. In fact, the prior probability density function penalizes the negative log-likelihood function $\ell_{protein}(\boldsymbol{y}|\theta)$. For the Gaussian prior $p_{Gauss}(\theta)$, we set $\theta_\iota^* = -1$ and $\sigma_{\theta_\iota}^2 = 100$ for every parameter in θ. A logarithmic function is defined for $\mathbb{R}^+ \setminus \{0\}$. This is the reason why we cannot set θ_ι^* to zero in the exponentially back-transformed parameter space. Nevertheless, deviations from zero are negligible regarding the used variances $\sigma_{\theta_\iota}^2$ in the non-transformed parameter space.

Optimization was done with the R package **dMod**[8] by using a method classified as trust region optimization which is implemented in the R function `trust()` [100, 108]. In **dMod**, model sensitivities $d_\theta \boldsymbol{C}(t, \theta)$ are calculated by solving the sensitivity equation

$$d_t d_\theta \boldsymbol{C}(t, \theta) = d_{\boldsymbol{C}(t,\theta)} \boldsymbol{f}(\boldsymbol{C}(t, \theta), \theta) \cdot d_\theta \boldsymbol{C}(t, \theta) + d_\theta \boldsymbol{f}(\boldsymbol{C}(t, \theta), \theta) \tag{5.24}$$

[8] freely available on `https://github.com/dkaschek/dMod`

jointly with the ODE system given in Equation 5.1 (see Appendix D.7 for calculation details). Therefore, functions $\boldsymbol{f}(\boldsymbol{C}(t,\theta),\theta))$ and $\boldsymbol{g}(\boldsymbol{C}(t,\theta),\theta))$ have to be continuously differentiable for variables $\boldsymbol{C}(t,\theta)$ and parameters θ [109]. We implemented transitions between low $k_{j,low}$ and high $k_{j,high}$ mRNA production rates by using Fermi-Dirac distributions (see Table 4.2 and Appendix D.8). The continuity of the observation function $\boldsymbol{g}(\boldsymbol{C}(t,\theta),\theta))$ is needed to calculate sensitivities of observables which are used in the gradient calculation. Dependent on the synchronization method, we changed timing in the simulation of Equation 5.1 and decided if initial values have to be estimated or not (see Figure E.3).

5.3 Combining count and time series data in a 2-step-optimization

The optimization problem for the chemical reaction system (see Table 4.2) is characterized by qualitatively different data. Further, smFISH and Western blot data have different error models. Combining different data types is not covered by available parameter estimation tools. We overcome this challenge by using a 2-step-optimization. First, we estimate mRNA parameters from smFISH data in the mRNA optimization step. Second, we re-estimate mRNA and estimate protein parameters from Western blot data in the protein optimization step.

Whereas the mRNA optimization step is completely separated (see Section 5.2), we make use of its results in the protein optimization step. We re-estimate mRNA parameters under the constraint to stay within the 95% confidence region of mRNA parameters estimated in the mRNA optimization step. Furthermore, we introduce a scaling factor to transcription start and end times to handle mismatches in cell division times between unsynchronized and synchronized cells. We have no information about phase lengths in Western blot data. This is why we can vary lengths and positions of high transcription regions only and equidistantly change phase lengths. However, deviating from a pure protein parameter estimation allows to successfully fit Western blot data for different synchronization methods.

As seen in Section 5.2, we can add another prior probability density function for mRNA parameters to the likelihood function in Equation 5.21 giving

$$L''_{protein}(\boldsymbol{y}|\theta) = L'_{protein}(\boldsymbol{y}|\theta) \cdot p_{mRNA}(\theta^r) \tag{5.25}$$

with the corresponding negative log-likelihood function

$$\ell''_{protein}(\boldsymbol{y}|\theta) = \ell'_{protein}(\boldsymbol{y}|\theta) + \sum_{\iota=1}^{nPar} \frac{1}{2}\left[\left(\frac{\theta_\iota^{r*} - \theta_\iota^r}{\sigma_{\theta_\iota^r}}\right)^2 + \log(2\pi\sigma_{\theta_\iota^r}^2)\right] \tag{5.26}$$

and

$$p_{mRNA}(\theta^r) = \prod_{\iota=1}^{nPar} \frac{1}{\sqrt{2\pi\sigma_{\theta_\iota^r}^2}} e^{-\frac{(\theta_\iota^{r*}-\theta_\iota^r)^2}{2\sigma_{\theta_\iota^r}^2}} \tag{5.27}$$

for $\theta_\iota^r \sim \mathcal{N}(\theta_\iota^{r*}, \sigma_{\theta_\iota^r}^2)$ where $\theta^r \subset \theta$ includes mRNA parameters only. We set known parameter values θ_ι^{r*} of the mRNA prior $p_{mRNA}(\theta^r)$ to its log-transformed parameter estimates of the mRNA optimization step and variances $\sigma_{\theta_\iota^r}^2$ to the log-transformed and subsequently squared distances between parameter estimate and border of its 95%-confidence region. A successful parameter estimation for hydroxyurea and nocodazole synchronized cells requires two times this distance. Whenever the 95% confidence region is not symmetric, we decided for the larger distance. In cases where the 95% confidence region is not uniquely determinable to one side, we decided for the determinable distance of the 95% confidence region.

5.4 Performing global optimization by multi-start local optimization

In rare cases, optimization processes are linear and convex meaning that the landscape of the objective function has an unique optimum and local (deterministic) optimization methods are

suited [99]. Optimization problems in our 2-step-optimization are non-linear and non-convex wherefore the landscape of the objective function has multiple local optima. For this reason, global (stochastic) optimization methods are required. Typical global optimization methods are multi-start local optimizations, genetic and evolutionary algorithms, particle swarm optimizations, simulated annealing, as well as hybrid optimizers [60].

We decided for a multi-start local optimization. The main idea of this method is to start several optimization runs with different initial guesses for parameters which are sampled from the multidimensional parameter space. Since each optimization run is independent, we can parallelize the process using the R function `mclapply()` of the R package **parallel**. In the mRNA optimization step, we used a latin hypercube sampling which is implemented in the R function `Latinhyper()` of the R package **FME**[9] (see Appendix D.10 for sampling details). We used the **dMod** function `mstrust()` to perform multi-start optimization in the protein optimization step which reverts to a Gaussian sampling. In general, we started 1000 to 2000 runs per optimization but not all of them end successfully. Reasons for a premature termination of an optimization are numerical errors in the calculation of the ODE solution due to the ODE solver or convergence failures in the optimization algorithm due to inaccuracies in the calculation of model sensitivities.

The performance of the optimization algorithm can be visualized in "Likelihood Waterfall" plots (see Figure 5.1). In these plots, values of the objective function are sorted in an in- or decreasing order. An efficient optimization method shows several clearly differentiable optima where the same value is reached a couple of times for different initial guesses. In case of minimizing the negative log-likelihood function, the lowest optimum is an indicator for the global optimum. Nevertheless, we can never be sure that we reached the global optimum because we cannot sample the whole parameter space.

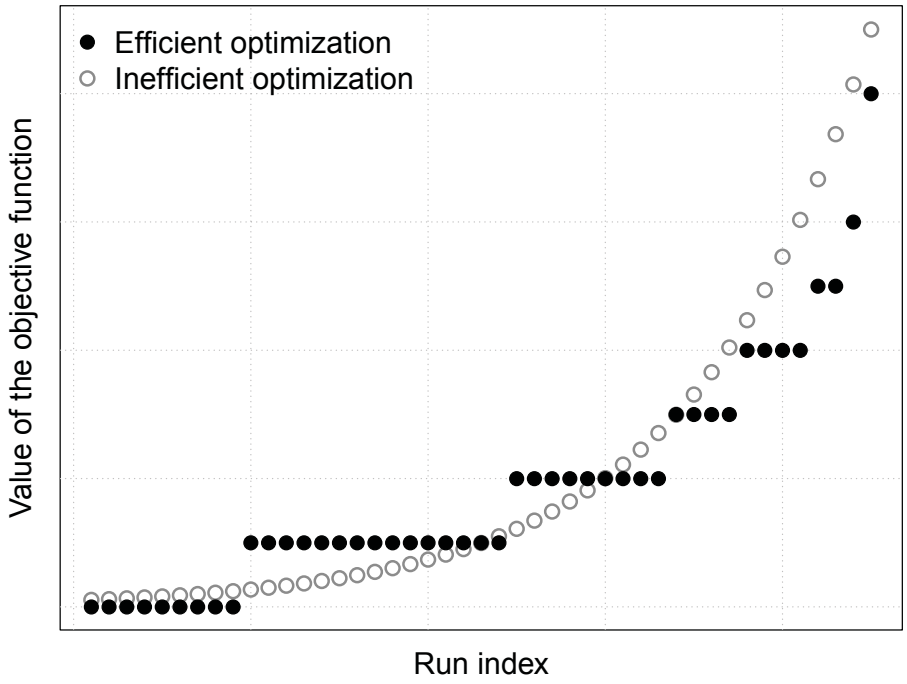

Figure 5.1: Visualizing the performance of a multi-start local optimization in a likelihood waterfall plot. In a multi-start local optimization, we start several optimization runs for different initial guesses of parameter values. A likelihood waterfall plot represents values of the objective function for each optimization run sorted by size. An efficient optimization method shows several clearly differentiable optima whereas an inefficient optimization method e.g. continuously increase. In the case of minimizing the negative log-likelihood function, the lowest optimum is an indicator of the global optimum.

[9] freely available on `https://cran.r-project.org/web/packages/FME`

6. Identifiability analysis

All definitions in this chapter are taken from [58, 104, 110, 111]. Different references are indicated.

6.1 Introducing the concept of parameter identifiability

Identifiability analysis is crucial in parameter estimation. Measurement uncertainties transfer via parameter uncertainties to prediction uncertainties. Mostly, experimental data are insufficient regarding the size of the chemical reaction system. At the same time, experimental data are noisy and variances are large. Following the law of large numbers, variances becomes smaller for an increasing number of replicates [61]. Regarding Western blot data, we only have a few replicates. A sufficient number of replicates is often not feasible due to elaborate experimental procedures or costs. On the mRNA level, we measured more than 900 cells per mRNA species but some cell cycle phases are less represented than others and variances can still be large. The interplay between the amount of experimental data, the measurement noise and the sampling rate determines the information content of the data [112].

Parameter uncertainties can lead to non-identifiabilities. Non-identifiable parameters have no unique solution [112]. We discriminate between practically and structurally non-identifiable parameters. Practical non-identifiabilities can be removed by improving amount and quality of the data, e.g. more replicates, additional time points, absolut measurements concerning Western blot data or more cells concerning smFISH data. In contrast, structural non-identifiabilities come from the structure of the chemical reaction system and can be eliminated by qualitatively new measurements which change the observation function $\boldsymbol{g}$ (see Equation 5.2 and Table 5.1).

Non-observabilities can result from prediction uncertainties. If trajectories of variable $C_i(t, \theta)$ are affected by a non-identifiable parameter and not determinable, this variable is non-observable. Trajectories of observables $\boldsymbol{Y}(t, \theta)$ are typically invariant against structural non-identifiabilities and vary for practical non-identifiabilities. As a consequence of non-observabilities, biological questions can, if at all, partially be answered.

We use the concept of profile likelihoods to investigate parameter identifiabilities and to determine parameter confidence intervals. Unlike asymptotic confidence intervals, likelihood based confidence intervals can detect asymmetric confidence intervals and, furthermore, detect non-identifiabilities. Asymptotic confidence intervals are calculated from the variance-covariance matrix as inverse of the Hessian of the objective function evaluated at the estimated parameter vector [63, 98]. These confidence intervals are symmetric and appropriate for identifiable parameters and highly informative data.

6.2 Working with profile likelihoods

Confidence intervals $[\varsigma^-, \varsigma^+]$ of parameter estimates $\hat{\theta}$ to a confidence level $(1 - \alpha)$ indicate that true values θ^* are located within these intervals with probability $(1 - \alpha)$. Likelihood based confidence intervals are defined by the likelihood ratio

$$\left\{ \theta \middle| \frac{L(\boldsymbol{y}|\theta)}{L(\boldsymbol{y}|\hat{\theta})} > c \right\} \tag{6.1}$$

with cutoff value $c = \chi^2_{df,(1-\alpha)}$ for a sufficiently large number of data points [98]. The cutoff value is the $(1-\alpha)$ quantile of the $\chi^2(df)$ distribution with the significance level α and the degree of freedom df. The most frequently used significance level is $\alpha = 0.05$. We decided to calculate pointwise confidence intervals with $df = 1$ that hold for each parameter individually. Using the

negative log-likelihood function, confidence intervals become

$$\{\theta|\ell(\boldsymbol{y}|\theta) - \ell(\boldsymbol{y}|\hat{\theta}) < -\log(c)\} \tag{6.2}$$

with cutoff value $c = e^{-\chi^2_{df,(1-\alpha)}}$.

The profile likelihood of the parameter $\theta_{\underline{\iota}}$ is defined as

$$PL(\boldsymbol{y}|\theta_{\underline{\iota}}) = arg \min_{\forall\theta_\iota \neq \theta_{\underline{\iota}}} \ell(\boldsymbol{y}|\theta) \tag{6.3}$$

and explore its parameter space along the least increase in the objective function. In this way, parameter $\theta_{\underline{\iota}}$ is consecutively fixed to a value in a certain range around its optimum $\hat{\theta}_{\underline{\iota}}$ and $\ell(\boldsymbol{y}|\theta)$ is re-optimized for remaining parameters $\theta_\iota \neq \theta_{\underline{\iota}}$. Thus, each re-optimization gives a different parameter vector $\hat{\theta}^{PL}$ [113]. We can now determine the confidence interval of each parameter by using the likelihood ratio

$$\{\theta_{\underline{\iota}}|PL(\boldsymbol{y}|\theta_{\underline{\iota}}) - PL(\boldsymbol{y}|\hat{\theta}_{\underline{\iota}}) < -log(c)\}. \tag{6.4}$$

Values $PL(\boldsymbol{y}|\theta_{\underline{\iota}})$ and $PL(\boldsymbol{y}|\hat{\theta}_{\underline{\iota}})$ correspond to values of the negative log-likelihood function $\ell(\boldsymbol{y}|\theta)$ evaluated at the re-optimized parameter vector $\hat{\theta}^{PL}$ and the parameter optimum $\hat{\theta}$, respectively. Likelihood ratios calculated on the basis of negative log-likelihood functions are represented by differences. In the following, we still name them ratios.

A parameter is identifiable if the confidence interval is finite $[\varsigma^-_{\underline{\iota}}, \varsigma^+_{\underline{\iota}}]$ meaning that the profile likelihood $PL(\boldsymbol{y}|\theta_{\underline{\iota}})$ crosses $\chi^2_{df,(1-\alpha)}$ for smaller and larger parameter values compared to its optimum (see Figure 6.1). A practically non-identifiable parameter has a one-sided infinite confidence interval and the profile likelihood $PL(\boldsymbol{y}|\theta_{\underline{\iota}})$ crosses $\chi^2_{df,(1-\alpha)}$ either for smaller $[\varsigma^-_{\underline{\iota}}, +\infty]$ or larger $[-\infty, \varsigma^+_{\underline{\iota}}]$ parameter values compared to its optimum. In contrast, a structurally non-identifiable parameter has an infinite confidence interval $[-\infty, +\infty]$ and the profile likelihood $PL(\boldsymbol{y}|\theta_{\underline{\iota}})$ never crosses the $\chi^2_{df,(1-\alpha)}$. The profile likelihood $PL(\boldsymbol{y}|\theta_{\underline{\iota}})$ is mostly constant and equal to the the profile likelihood $PL(\boldsymbol{y}|\hat{\theta}_{\underline{\iota}})$ in infinite confidence regions why their difference is zero.

Variability of variables $\boldsymbol{C}(t,\theta)$ and observables $\boldsymbol{Y}(t,\theta)$ can be analyzed by plotting trajectories for parameter values along the profile likelihood. Trajectories reveal regions where parameter uncertainty has the largest influence on the chemical reaction system. Parameter dependencies can be analyzed by plotting parameter values θ_ι against parameter values of the profiled parameter $\theta_{\underline{\iota}}$ (see Figure 6.2). Flat lines indicate independent parameters ($|var(\theta)| < 0.01$) whereas in- or decreasing lines show dependent parameters ($|var(\theta)| > 0.01$).

We used the R function `profile()` of the R package **dMod** to calculate profile likelihoods in the mRNA and the protein optimization step.

mRNA optimization step

According to Section 5.2, the profile likelihood of the mRNA optimization step becomes

$$PL_{mRNA}(y_k|\theta^k_{\underline{\iota}}) = arg \min_{\forall\theta^k_\iota \neq \theta^k_{\underline{\iota}}} \ell_{mRNA}(y_k|\theta^k) \tag{6.5}$$

and the corresponding likelihood based confidence interval is

$$\{\theta^k_{\underline{\iota}}|PL_{mRNA}(y_k|\theta^k_{\underline{\iota}}) - PL_{mRNA}(y_k|\hat{\theta}^k_{\underline{\iota}}) < -log(c)\}. \tag{6.6}$$

Total profiles of the mRNA optimization step have a single contribution. It is the contribution of the data feeding in the negative log-likelihood function $\ell_{mRNA}(y_k|\theta^k)$. As described above, non-identifiabilities can be read directly from profile likelihoods (see Figure 6.1).

Protein optimization step

According to Section 5.3, the profile likelihood of the protein optimization step becomes

$$PL_{protein}(\boldsymbol{y}|\theta_{\underline{\iota}}) = arg \min_{\forall\theta_\iota \neq \theta_{\underline{\iota}}} \ell''_{protein}(\boldsymbol{y}|\theta) \tag{6.7}$$

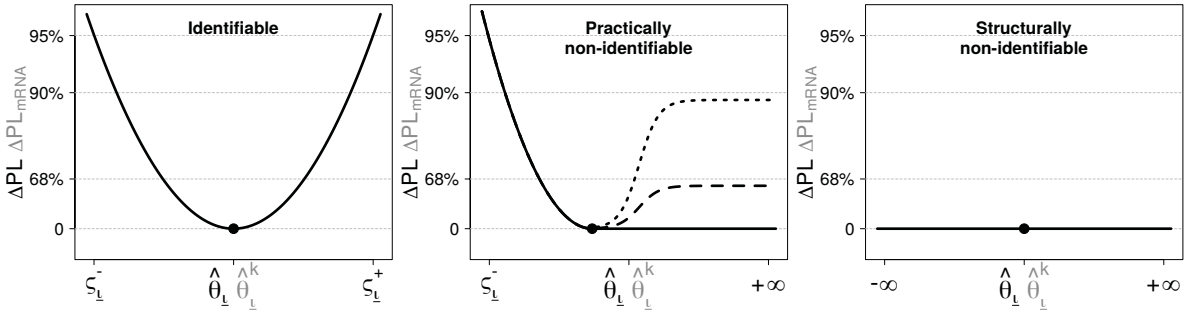

Figure 6.1: Calculating confidence intervals and determining identifiabilities by using profile likelihoods. In this figure, we show how profile likelihoods are used to determine confidence intervals and identifiabilies of parameter estimates. The profile likelihood $PL(\boldsymbol{y}|\theta_{\underline{\iota}})$ of a parameter $\theta_{\underline{\iota}}$ defined in Equation 6.3 explores its parameter space along the least increase in the objective function $\ell(\boldsymbol{y}|\theta)$. We use the profile likelihood to calculate the likelihood based confidence interval $\{\theta_{\underline{\iota}}|PL(\boldsymbol{y}|\theta_{\underline{\iota}}) - PL(\boldsymbol{y}|\hat{\theta}_{\underline{\iota}}) < -log(c)\}$ of parameter $\theta_{\underline{\iota}}$ with cutoff value $c = e^{-\chi^2_{df,(1-\alpha)}}$. Plotting likelihood ratios $\Delta PL = PL(\boldsymbol{y}|\theta_{\underline{\iota}}) - PL(\boldsymbol{y}|\hat{\theta}_{\underline{\iota}})$ against parameter values $\theta_{\underline{\iota}}$ enables to determine confidence intervals (black labels). Respective confidence levels are given on the y-axis. We marked borders of the 95% confidence interval for illustration purposes. The confidence interval of an identifiable parameter is finite $[\varsigma_{\underline{\iota}}^-, \varsigma_{\underline{\iota}}^+]$. A practically non-identifiable parameter has a one-sided infinite confidence interval $[-\infty, \varsigma_{\underline{\iota}}^+]$ or $[\varsigma_{\underline{\iota}}^-, +\infty]$. In the given example, different scenarios of $[\varsigma_{\underline{\iota}}^-, +\infty]$ are illustrated. A structurally non-identifiable parameter has an infinite confidence interval $[-\infty-, +\infty]$. Each point in the profile corresponds to a different parameter vector $\hat{\theta}^{PL}$ [113]. Profile likelihoods of the mRNA optimization step $PL_{mRNA}(y_k|\theta_{\underline{\iota}}^k)$ defined in Equation 6.5 are represented in the same way. The total profile has a single contribution from the data. The likelihood ratio turns into $\Delta PL_{mRNA} = PL_{mRNA}(y_k|\theta_{\underline{\iota}}^k) - PL_{mRNA}(y_k|\hat{\theta}_{\underline{\iota}}^k)$ and the x-axis shows parameter $\theta_{\underline{\iota}}^k$ (gray labels).

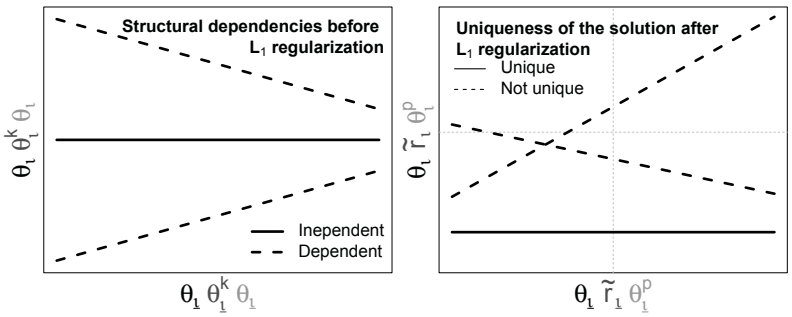

Figure 6.2: Parameter values along the profile likelihood. In this figure, we plot parameter values θ_{ι} against parameter values of the profiled parameter $\theta_{\underline{\iota}}$. Parameter dependencies become visible in profile likelihoods which are calculated before applying L_1 regularization (left panel) by using the negative log-likelihood function $\ell(\boldsymbol{y}|\theta)$. A flat line (solid line, $|var(\theta)| < 0.01$) indicates independent parameters whereas an in- or decreasing line (dashed line, $|var(\theta)| > 0.01$) shows that parameter θ_{ι} and profiled parameter $\theta_{\underline{\iota}}$ are dependent. Specific notations for the mRNA and the protein optimization step are given in gray and light gray labels, respectively. After applying L_1 regularization (right panel), we use profile likelihoods to test uniqueness of selected parameters. Therefore, we use the η specific negative log-likelihood function $\ell_\eta(\boldsymbol{y}|\theta^\eta)$ and increase the degree of freedom by introducing a non-selected parameter. The profile likelihood of the introduced parameter $\theta_{\underline{\iota}}$ is located around zero. Here, a flat line indicates that the introduced parameter was uniquely set to zero. If a line horizontally crosses zero, parameter θ_{ι} can be set to zero instead of the introduced parameter. Zero values are given by gray dashed lines. Specific notations for mRNA fold changes and protein parameters are given as gray and light gray labels, respectively.

and the corresponding likelihood based confidence interval is

$$\{\theta_{\underline{\iota}}|PL_{protein}(\boldsymbol{y}|\theta_{\underline{\iota}}) - PL_{protein}(\boldsymbol{y}|\hat{\theta}_{\underline{\iota}}) < -log(c)\}. \tag{6.8}$$

Total profiles of the protein optimization step have three contributions: data $\ell_{protein}(\boldsymbol{y}|\theta)$, mRNA prior $p_{mRNA}(\theta^r)$ and Gaussian prior $p_{Gauss}(\theta)$ (see Figure 6.3). All contributions feed in the negative log-likelihood function $\ell''_{protein}(\boldsymbol{y}|\theta)$. The usage of a Gaussian prior always leads to a crossing between total profile and $\chi^2_{df,(1-\alpha)}$ but corresponding parameter values become very small or very large. Non-identifiabilities become visible if the total profile follows the Gaussian prior either to one side (practically non-identifiable) or to both sides (structurally non-identifiable). In practice, we determine parameter identifiability with respect to a maximum permitted deviation from the parameter optimum depending on the $\chi^2_{df,(1-\alpha)}$ and the used synchronization method.

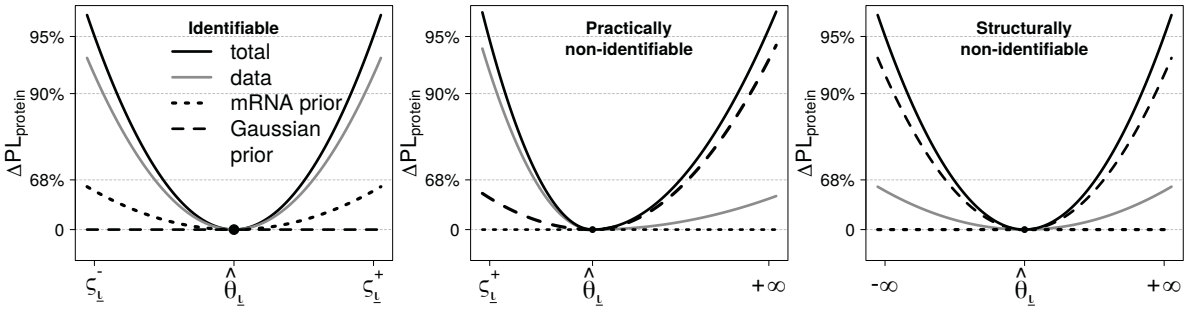

Figure 6.3: Profile likelihoods of the protein optimization step. In this figure, we shows profile likelihoods of the protein optimization step. The profile likelihood $PL_{protein}(\boldsymbol{y}|\theta_{\underline{\iota}})$ defined in Equation 6.7 has three contributions which determine the total profile (black solid line): data $\ell_{protein}(\boldsymbol{y}|\theta)$ (gray solid line), mRNA prior $p_{mRNA}(\theta^r)$ (black dotted line) and Gaussian prior $p_{Gauss}(\theta)$ (balck dashed line). Likelihood ratios $\Delta PL_{protein} = PL_{protein}(\boldsymbol{y}|\theta_{\underline{\iota}}) - PL_{protein}(\boldsymbol{y}|\hat{\theta}_{\underline{\iota}})$ are shown on the y-axis and parameter values $\theta_{\underline{\iota}}$ on the x-axis. Confidence intervals $\{\theta_{\underline{\iota}}|PL_{protein}(\boldsymbol{y}|\theta_{\underline{\iota}}) - PL_{protein}(\boldsymbol{y}|\hat{\theta}_{\underline{\iota}}) < -log(c)\}$ for different cutoff values $c = e^{-\chi^2_{df,(1-\alpha)}}$ are given on the y-axis. The 95% confidence interval is marked by $[\varsigma^-_{\underline{\iota}}, \varsigma^+_{\underline{\iota}}]$. The Gaussian prior causes a total profile which always exceed the 95% confidence level even if the parameter is either practically or structurally non-identifiable. Non-identifiabilities are characterized by a total profile following the Gaussian prior. Confidence intervals of practically non-identifiable parameters are infinite to one side, here exemplified for $[\varsigma^-_{\underline{\iota}}, +\infty]$. Structurally non-identifiable parameters have infinite confidence intervals to both sides $[-\infty, +\infty]$. The Gaussian prior prevent an actually infinite border. Nevertheless, values become very large. Each point in the total profile corresponds to a different parameter vector $\hat{\theta}^{PL}$ [113].

6.3 Classification of protein profile effects

Regarding the protein optimization step, we have seen in Section 5.2 and 6.2 that the total profile has three contributions. The Gaussian prior should not contribute to the total profile of any parameter. If the Gaussian prior still determine the total profile, the parameter is non-identifiable.

The mRNA prior was introduced to combine the mRNA and protein optimization step. This mRNA prior allows for small parameter re-adjustments to overcome differences between unsynchronized and synchronized cells. The contribution of the mRNA prior should only affect re-estimated mRNA parameters. Re-estimated mRNA parameters whose total profiles are only determined by the mRNA prior are not affected by Western blot data and point to shared parameter values among unsynchronized and synchronized cells. If the mRNA prior contributes to one side of the total profile of a protein parameter, Western blot data cannot fully determine the parameter value. In a different scenario, the total profile of a protein parameter can be completely determined by the mRNA prior. In this case, Western blot data do not include any information about the parameter value.

Similarly, we only expect a contribution of the Western blot data to protein parameters. Western blot data are fully informative if the total profile of a protein parameter is only determined by the contribution of the data. A contribution of the data to either one or two sides of the total profile of a re-estimated mRNA parameter indicates a mismatch between unsynchronized and synchronized cells. Therefore, it is important to classify the contribution of the mRNA prior to the total profile of protein parameters or the contribution of the data to the total profile of re-estimated mRNA parameters.

A special case is the scaling factor introduced to transcription start and end times in the protein optimization step to combine mRNA and protein optimization step as well. This parameter cannot clearly be assigned to mRNA or protein parameters. Its value is dependent on Western blot data and scales mRNA parameters. A contribution of the mRNA prior is equally likely as a contribution of the data.

We classify mRNA prior or data contributions to total profiles depending on which side of the parameter optimum the respective contribution occurs and if this effect is dominant. Effects can occur on the left-hand side (LHS) for smaller parameter values, on the right-hand side (RHS) for larger parameter values or on both sides (LRHS) of its optimum. Satisfying a certain criteria, an effect can be strong, weak or not relevant. A contribution can switch from not relevant to weak by increasing the parameter region used for the classification. This problem cannot be solved if a contribution has an increasing tendency. Only strong effects are unaffected by an increased parameter region. This is why we decided to use the parameter range covered by the total profile for the classification. We used the following rules (see Figure 6.4):

1. No effect: mRNA prior or data contribution stays below the 68% confidence level
2. Weak left effect: mRNA prior or data contribution is smaller than the total profile but exceed the 68% confidence level for smaller parameter values compared to the optimum
3. Strong left effect: mRNA prior or data contribution is larger than the total profile for smaller parameter values compared to the optimum
4. Weak right effect: mRNA prior or data contribution is smaller than the total profile but exceed the 68% confidence level for larger parameter values compared to the optimum
5. Strong right effect: mRNA prior or data contribution is larger than the total profile for larger parameter values compared to the optimum
6. Weak left and weak right effect: mixture of 2. and 4.
7. Weak left and strong right effect: mixture of 2. and 5.
8. Strong left and weak right effect: mixture of 3. and 4..

It is not possible to have a strong effect on both sides.

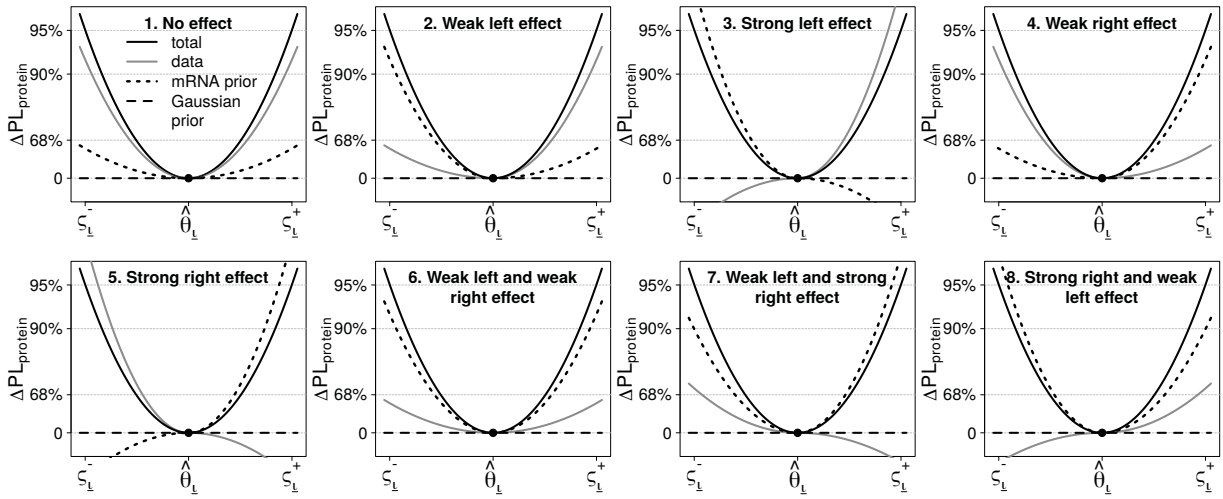

Figure 6.4: Classification of mRNA prior contributions to the total profile of a protein parameter. In this figure, we illustrate the classification for mRNA prior contributions to the total profile of a protein parameter. The profile likelihood $PL_{protein}(\boldsymbol{y}|\theta_{\underline{\iota}})$ defined in Equation 6.7 of the protein optimization step has three contributions to the total profile (black solid line): data $\ell_{protein}(\boldsymbol{y}|\theta)$ (gray solid line), mRNA prior $p_{mRNA}(\theta^r)$ (black dotted line) and Gaussian prior $p_{Gauss}(\theta)$ (black dashed line). Each diagram plots the likelihood ratio $\Delta PL_{protein} = PL_{protein}(\boldsymbol{y}|\theta_{\underline{\iota}}) - PL_{protein}(\boldsymbol{y}|\hat{\theta}_{\underline{\iota}})$ against the parameter value $\theta_{\underline{\iota}}$. 95% confidence intervals $[\varsigma_{\underline{\iota}}^-, \varsigma_{\underline{\iota}}^+]$ are marked. The mRNA prior has no effect if its contribution stays below the 68% confidence level. In a weak contribution (LHS or RHS), the mRNA prior contribution is smaller than the total profile and exceed the 68% confidence level. A strong contribution (LHS or RHS) means that the mRNA prior is larger than the total profile. Mixtures between weak and strong contributions can occur (LRHS), but a strong contribution to both sides is not possible.

7. L_1 regularization

All definitions in this chapter are taken from [114, 115]. Different references are indicated.

7.1 Regularizing the likelihood function

A common problem in parameter estimation are non-identifiable parameters which arise from the structure of the chemical reaction system or from insufficient information in the data. Whereas structural non-identifiabilities can be removed by qualitatively new measurements, practical non-identifiabilities disappear if e.g. additional data points are measured. Moreover, parameter estimation can be improved by using regularization methods leading to model reduction.

Adding a regularization term to the negative log-likelihood function penalizes parameters not covered by the data and its values go to zero in the optimization procedure. The penalized negative log-likelihood function become

$$\ell_{L_\rho}(\boldsymbol{y}|\theta) = \ell(\boldsymbol{y}|\theta) + \eta\|\theta\|_\rho \tag{7.1}$$

with the regularization strength η and the ρ-norm of the parameter vector θ

$$\|\theta\|_\rho = \left(\sum_{\iota=1}^{nPar} |\theta_\iota|^\rho\right)^{1/\rho}. \tag{7.2}$$

By setting $\rho = 1$, we get the L_1 regularized negative log-likelihood function

$$\ell_{L_1}(\boldsymbol{y}|\theta) = \ell(\boldsymbol{y}|\theta) + \eta \sum_{\iota=1}^{nPar} |\theta_\iota|. \tag{7.3}$$

The regularization strength η controls how many parameters remain different from zero and are finally selected in the optimization. The number of non-selected parameters increase with an increasing regularization strength. The most challenging step is to determine the optimal regularization strength $\hat{\eta}$.

Optimizing Equation 7.3 gives biased parameter estimates $\hat{\theta}^{bias}$ with $nPar_0$ zero valued parameters. In contrast, optimizing a η specific negative log-likelihood function $\ell_\eta(\boldsymbol{y}|\theta^\eta)$ gives constrained but unbiased parameter estimates $\hat{\theta}^\eta$. Parameter vector θ^η includes $nPar - nPar_0$ parameters which are determined as being non-zero in the respective biased parameter estimate $\hat{\theta}^{bias}$. Other parameters are fixed to zero.

We systematically test different regularization strengths and use the likelihood ratio

$$D(\eta) = \ell_\eta(\boldsymbol{y}|\hat{\theta}^\eta) - \left[\ell_\eta(\boldsymbol{y}|\hat{\theta}^\eta)\right]_{\eta=0} \tag{7.4}$$

to maximize η satisfying

$$D(\eta) < \chi^2_{df,(1-\alpha)} \tag{7.5}$$

with $df = nPar - nPar_0$. As long as the criterion is fulfilled, the reduced model is as good as the full model. Basically, we test the null hypothesis that there is no significant difference between reduced and full model.

By using the profile likelihood

$$PL_\eta(\boldsymbol{y}|\theta^\eta_{\underline{\iota}}) = arg \min_{\forall\theta^\eta_\iota \neq \theta^\eta_{\underline{\iota}}} \ell_\eta(\boldsymbol{y}|\theta^\eta), \tag{7.6}$$

we test if we successfully selected all parameters covered by the data. Parameter values should not be compatible with zero meaning that the profiled parameter does not become zero within the 95% confidence region. If the profiled parameter still becomes zero, this parameter was false-positively selected and we can subsequently set its value to zero which is called supervised removal (see Figure 7.1).

Further, we use the profile likelihood to test uniqueness of selected parameters. We increase the degree of freedom in the η specific negative log-likelihood function $\ell_\eta(\boldsymbol{y}|\theta^\eta)$ by introducing a non-selected parameter. Since this parameter was set to zero during the L_1 regularization, its profile likelihood is located around zero. Plotting all parameter values θ_ι against the profiled parameter $\theta_{\underline{\iota}}$ reveal the uniqueness of the solution (see Figure 6.2). A parameter shows a flat line if the introduced parameter is uniquely determined to be zero. Differently, a parameter which becomes compartible with zero after introducing a non-selected parameter shows shows a line which horizontally crosses zero. In this case, the introduced parameter can deviate from zero and the opposing parameter becomes zero instead. Consequently, we cannot decide which parameter should be selected.

We applied L_1 regularization in the protein optimization step and used the R function `L1trust()` of the R package **dMod** to determine the biased parameter estimate $\hat{\theta}^{bias}$. In this optimization step, all parameter values are log-transformed. That means that a zero valued parameter after applying L_1 regularization is set to one in a non-transformed parameter space. Since we do not expect information about mRNA parameters in the Western blot data, we cannot use a common regularization for mRNA and protein parameters and discriminate between them. How we apply L_1 regularization for both parameter subsets is explained in Sections 7.2 and 7.3.

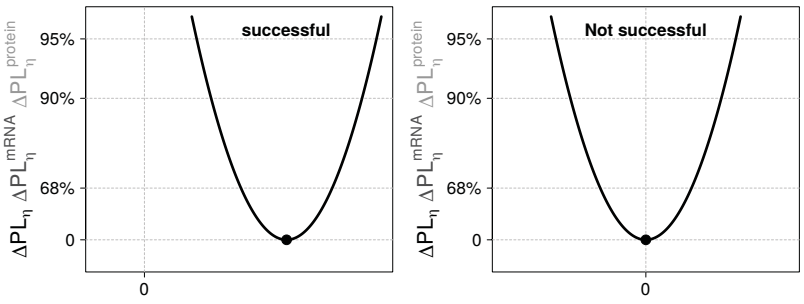

Figure 7.1: Success of parameter selection after L_1 regularization. In this figure, we show how to use the profile likelihood $PL_\eta(\boldsymbol{y}|\theta^\eta_{\underline{\iota}})$ defined in Equation 7.6 to test if we successfully selected all parameters covered by the data. The likelihood ratio $\Delta PL_\eta = PL(\boldsymbol{y}|\theta^\eta_{\underline{\iota}}) - PL(\boldsymbol{y}|\hat{\theta}^\eta_{\underline{\iota}})$ is given on the y-axis with relevant confidence levels and parameter values $\theta^\eta_{\underline{\iota}}$ on the x-axis. The focus of the illustration is on the x-position of zero relative to the parameter optimum (point) why no other values are marked. The profile likelihood of a successful selected parameter is not compatible with zero (left panel) within the 95% confidence region whereas it is for a false-positively selected parameter (right panel). Notations specific for mRNA fold changes and protein parameters are given as gray and light gray labels, respectively.

7.2 Identifying protein parameters not covered by the data

We apply L_1 regularization in the protein optimization step to protein parameters $\theta^p \subset \theta$ to figure out which parameters are not determinable by the given data. mRNA parameters are fixed to its estimates resulting from optimizing $\ell''_{protein}(\boldsymbol{y}|\theta)$ (see Equation 5.26). For the fixation it is required that mRNA parameters are identifiable. If this requirement is not fulfilled, we cannot reliably judge protein parameters. The regularized negative log-likelihood function becomes

$$\ell_{L_1}^{protein}(\boldsymbol{y}|\theta^p) = \ell'_{protein}(\boldsymbol{y}|\theta^p) + \eta \sum_{\iota=1}^{nPar} |\theta^p_\iota|. \tag{7.7}$$

The likelihood ratio reads

$$D_{protein}(\eta) = \ell_{\eta}^{protein}(\boldsymbol{y}|\hat{\theta}^{p,\eta}) - \left[\ell_{\eta}^{protein}(\boldsymbol{y}|\hat{\theta}^{p,\eta})\right]_{\eta=0} \tag{7.8}$$

where $\ell_{\eta}^{protein}(\boldsymbol{y}|\tilde{r}^{\eta})$ is the η specific negative log-likelihood function. The profile likelihood is

$$PL_{\eta}^{protein}(\boldsymbol{y}|\theta_{\underline{\iota}}^{p,\eta}) = arg \min_{\forall\theta_{\iota}^{p,\eta}\neq\theta_{\underline{\iota}}^{p,\eta}} \ell_{\eta}^{protein}(\boldsymbol{y}|\theta^{p,\eta}). \tag{7.9}$$

In the best possible case, structurally non-identifiable parameters go to zero and practically non-identifiable parameters become identifiable [116].

Since we do not estimate mRNA parameters, the regularized negative log-likelihood function has no mRNA prior contribution anymore. The scaling factor introduced to the transcription start and end times is fixed as well.

7.3 Calculating fold changes between mRNA parameters of unsynchronized and synchronized cells

We apply L_1 regularization in the protein optimization step to mRNA parameters θ^r to figure out which mRNA parameters of synchronized cells are different from mRNA parameters of unsynchronized cells. Parameter θ_{ι}^r is defined as $\theta_{\iota}^r = \tilde{r}_{\iota}\theta_{\iota}^{FISH}$ with θ_{ι}^{FISH} the ι-th parameter of the parameter vector $\theta^{FISH} = (\theta^{k=1}, \theta^{k=3}, \theta^{k=5})$ including all mRNA parameters which are estimated in the mRNA optimization step and log-transformed mRNA fold change $\tilde{r}_{\iota}$. The regularized negative log-likelihood function becomes

$$\ell_{L_1}^{mRNA}(\boldsymbol{y}|\tilde{r}) = \ell'_{protein}(\boldsymbol{y}|\tilde{r}) + \eta \sum_{\iota=1}^{nPar} |\tilde{r}_{\iota}|. \tag{7.10}$$

The likelihood ratio reads

$$D_{mRNA}(\eta) = \ell_{\eta}^{mRNA}(\boldsymbol{y}|\hat{\tilde{r}}^{\eta}) - \left[\ell_{\eta}^{mRNA}(\boldsymbol{y}|\hat{\tilde{r}}^{\eta})\right]_{\eta=0} \tag{7.11}$$

where $\ell_{\eta}^{mRNA}(\boldsymbol{y}|\tilde{r}^{\eta})$ is the η specific negative log-likelihood function. The profile likelihood is

$$PL_{\eta}^{mRNA}(\boldsymbol{y}|\tilde{r}_{\underline{\iota}}^{\eta}) = arg \min_{\forall\tilde{r}_{\iota}^{\eta}\neq\tilde{r}_{\underline{\iota}}^{\eta}} \ell_{\eta}^{mRNA}(\boldsymbol{y}|\tilde{r}^{\eta}). \tag{7.12}$$

mRNA fold change $\tilde{r}_{\iota}$ is going to zero during the optimization procedure if mRNA parameters are equal in unsynchronized and synchronized cells. Thus, a non-zero mRNA fold change $\tilde{r}_{\iota}$ indicates different mRNA parameters. In the non-transformed parameter space, $\tilde{r}_{\iota} = 1$ for equal, $\tilde{r}_{\iota} < 1$ for smaller or $\tilde{r}_{\iota} > 1$ for larger mRNA parameters. In the best possible case, calculated mRNA fold changes coincide with data contributions to the total profile of mRNA parameters described in Section 6.3 and illustrated in Figure 6.4.

Protein parameters are fixed to its L_1 regularized parameter estimates resulting from optimizing $\ell_{L_1}^{protein}(\boldsymbol{y}|\theta^p)$ (see Equation 7.7). We cannot use parameter estimates from optimizing $\ell''_{protein}(\boldsymbol{y}|\theta)$ (see Equation 5.26) because some protein parameters will be non-identifiable. It is also required that mRNA parameters θ^{FISH} of the mRNA optimization step are identifiable. Otherwise, calculating mRNA fold changes makes no sense at all.

The regularized negative log-likelihood function has no mRNA prior contribution. The mRNA prior is for mRNA parameters only and not for mRNA fold changes. The scaling factor introduced to the transcription start and end times is fixed as well.

III Results

8 Experimental data 63
8.1 smFISH data reveal low numbers of mRNA molecules
8.2 Western blot data show differences in protein abundances between synchronization methods
8.3 Pre-processing indicates synchronization specific protein numbers

9 Mathematical modeling 67
9.1 Synchronization influences cell cycle timing and gene transcription
9.2 Observables successfully reproduce protein time courses of different synchronization methods

10 Parameter estimation 71
10.1 mRNA optimization perform superior to protein optimization
10.2 Parameters related to cell cycle timing are most clearly affected by synchronization

11 Identifiability analysis 77
11.1 mRNA parameters are generally identifiable
11.2 Cell cycle timing and gene transcription are equally affected by Western blot data
11.3 Protein parameters are predominantly practically non-identifiable
11.4 mRNA priors contribute less to protein parameters of α-factor synchronized cells
11.5 Parameter dependencies persist between identifiable and non-identifiable parameters
11.6 Variations in model trajectories occur equally for identifiable and non-identifiable parameters

12 L_1 regularization 89
12.1 L_1 regularization improves identifiability of protein parameters
12.2 mRNA fold changes suggest synchronization specific smFISH measurements

8. Experimental data

8.1 smFISH data reveal low numbers of mRNA molecules

We measured absolut numbers of mRNA molecules for mRNA species *SIC1*, *CLN2* and *CLB5* in unsynchronized single cells by smFISH (see Section 3.1). As a result of assigning each cell to a specific cell cycle phase by morphological markers, we got phase-resolved mRNA distributions (see Figures 8.1). The cell division time is 128.66 ± 10.66 minutes (mean $\pm$ SD). Individual cell cycle phase lengths are given in Table 8.1.

mRNA distributions are diverse. *CLB5* shows less than half of the mRNA numbers measured for *SIC1* and *CLN2*. In some mRNA distributions, the counted number of zero mRNA molecules is largest, especially for *CLN2*. *SIC1* has an atypical distribution in ana- and T/C phase and *CLN2* in late G1 phase. As a whole, low numbers are more frequent than large numbers. Even if we assume Poisson distributed mRNA distributions and there are mRNA distributions looking similar to a single or a two component Poisson distribution, only a few can really be represented by a Poisson distribution (see Figures E.4, E.5 and E.6). With the exception of *SIC1* and *CLN2* in P/M phase, the χ^2 goodness of fit test indicates that mRNA distributions are significantly different to a Poisson distribution (p-values < 0.05, see Appendix D.9 for calculation details).

8.2 Western blot data show differences in protein abundances between synchronization methods

We measured relative numbers of protein molecules for protein species Sic1, Cln2 and Clb5 of synchronized cell populations by Western blotting (see Section 3.2). Protein abundances are given as time courses over the cell cycle (see Figure 8.2). Different synchronization methods have different effects on cell cycle events. In physical synchronized cell populations (elutriation), Sic1 has the largest protein level. In contrast, Cln2 is largest in chemical synchronized cell populations (α-factor, hydroxyurea and nocodazole). Clb5 presents the smallest protein level in every synchronization.

Elutriated G1 daughter cells start growing before entering S phase why cells stay longer in G1 phase at an approximately constant protein level (first 50 to 70 minutes). This is why the cell division time (150 minutes) is longer compared to unsynchronized (see Table 8.1) or chemical synchronized cells (see Table 8.2). α-factor synchronizes cells in G1 phase but cells quickly enter

Cell cycle phase	Measurement [min]	Simulation [min]
Early G1	27.2 ± 1.4	27
Late G1	33.7 ± 2.4	34
S	20.5 ± 1.8	21
G2	21.4 ± 1.7	22
P/M	9.8 ± 0.6	10
Ana	10.4 ± 1.9	11
T/C	4.1 ± 1.0	4

Table 8.1: Cell cycle phase lengths in unsynchronized cells. This table presents individual cell cycle phase lengths in minutes which are calculated from up to four biological replicates and are given as mean $\pm$ SD (second column). Cell cycle phase lengths are calculated from the number of cells per cell cycle phase which is directly proportional to its length [72]. We distinguished between seven phases (first column): early G1, late G1, S, G2, pro-/metaphase (P/M), anaphase (Ana) and telophase/cytokinesis (T/C). The cell division time is 128.66 ± 10.66 minutes (mean $\pm$ SD). In simulations, we used slightly different integer valued phase lengths which sum up to 129 minutes (third column).

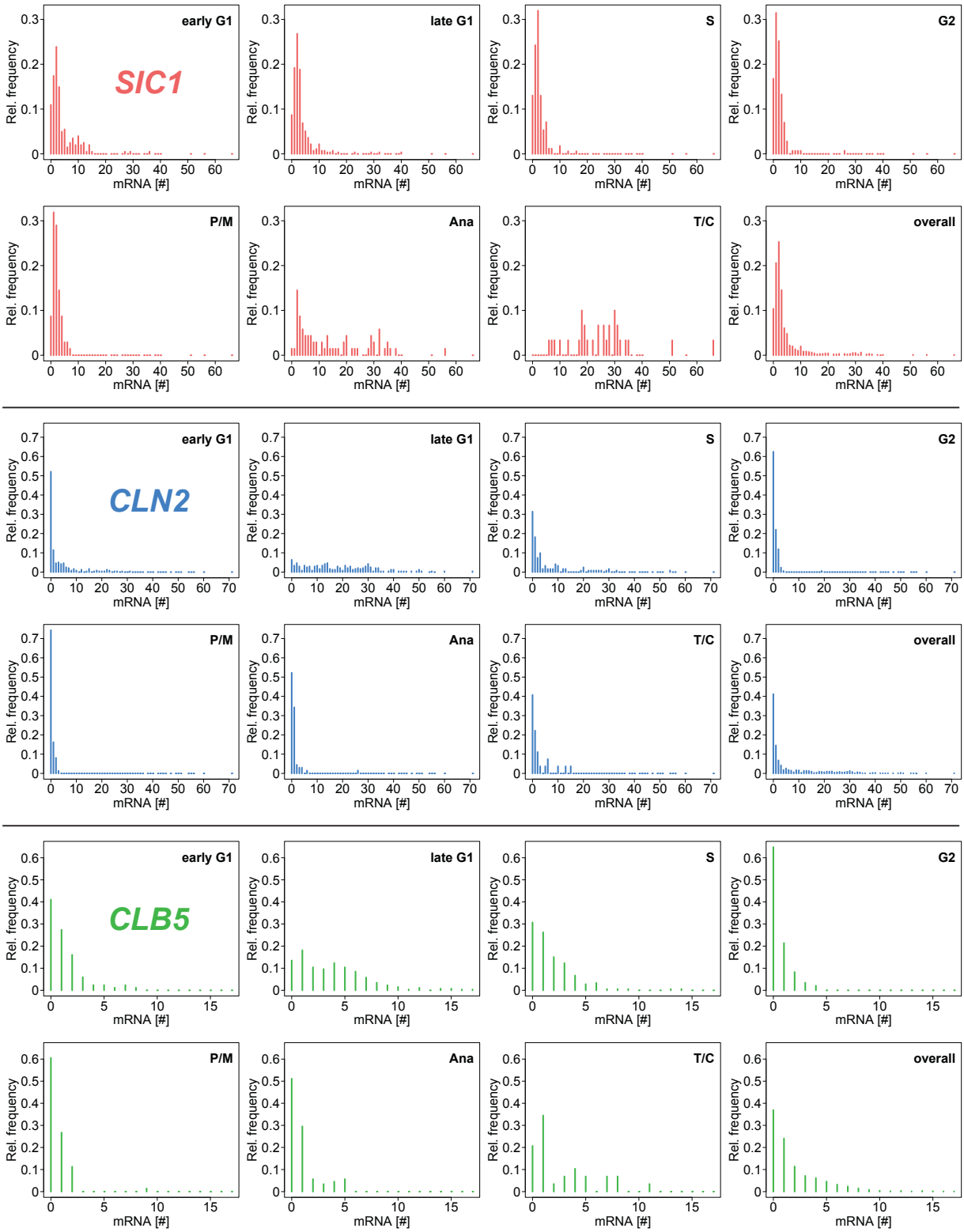

Figure 8.1: smFISH data of mRNA species *SIC1*, *CLN2* and *CLB5*. In this figure, phase-resolved mRNA distributions of *SIC1* (red), *CLN2* (blue) and *CLB5* (green) are shown. For reasons of comparability, we plotted relative frequencies against the number of mRNA molecules. We distinguished between seven cell cycle phases: early G1, late G1, S, G2, pro-/metaphase (P/M), anaphase (Ana) and telophase/cytokinesis (T/C). Additionally, we plotted the mRNA distribution over the whole cell cycle. This figure shows mRNA counts of 957, 957 and 928 cells in total for *SIC1*, *CLN2* and *CLB5*, respectively.

Feature	Elutriation	α-factor	Hydroxyurea	Nocodazole
Synchronization phase	G1	G1	S	G2
Cell division time	150	70	70	60
Replicates	4	2	2	4
Experiments	2	1	1	2

Table 8.2: Characteristics of synchronization specific Western blot data. This table shows synchronization phases, cell division times, numbers of technical replicates and numbers of experiments for each synchronization method.

into S phase. Compared to protein time courses of elutriated cells, Sic1 decrease immediately after release as well as Cln2 and Clb5 increase.

Nocodazole and hydroxyurea synchronize cells in later cell cycle phases. Proteins peak considerabely later in the cell cycle compared to α-factor synchronized cells. Cells have a longer S phase after synchronization by hydroxyurea and a short G2 phase after synchronization by nocodazole. Thus, protein peaks of hydroxyurea synchronized cells occur later in the cell cycle compared to nocodazole synchronized cells. Usually, the maximal protein level of Cln2 and Clb5 is close together. This behavior is disrubted in the first cell cycle passage of hydroxyurea synchronized cells.

Protein oscillations disappear as desynchronization progresses. As seen in protein time courses of elutriated cells, synchronization is weak. Oscillations already disappear after the first cell cycle passage. In contrast, nocodazole strongly synchronizes cells and ocsillations persist over three cell cycle passages.

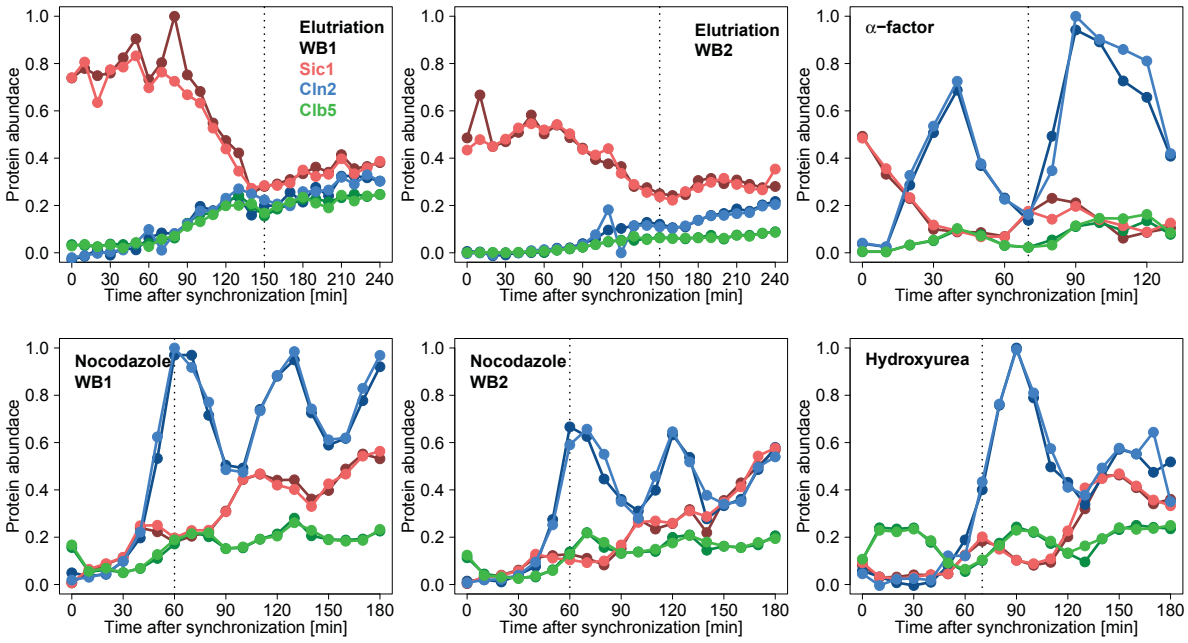

Figure 8.2: Western blot data of protein species Sic1, Cln2 and Clb5. In this figure, we represent protein time courses of elutriated, α-factor synchronized, hydroxyurea synchronized and nocodazole synchronized cell populations. Each panel shows a measurement with two applications per gel (light and dark colored). Different experiments are indicated by WB1 and WB2. Proteins are colored as follows: Sic1 in red, Cln2 in blue and Clb5 in green. We measured each protein with a sampling rate of ten minutes. Cell division times differ between synchronizations and are marked by black dotted lines. All data sets are normalized to its maximum whereby measurements of the same synchronization method have a common maximum. Differences between protein levels per synchronization method are real differences and not caused by normalization. Lines between data points are for visualization only.

8.3 Pre-processing indicates synchronization specific protein numbers

We used the R package **blotIt2** and the PaxDb database to pre-process (see Section 3.3) and subsequently normalize Western blot data (see Section 3.4). Since we want to study how synchronization affects the cell cycle behavior, we consider the first cell cycle passage only. Depending on the number of technical replicates (see Table 8.2) and variability between time courses, errors become smaller or larger (see Figure 8.3). Errors are smallest for α-factor and nocodazole synchronized cells and largest for elutriated cells, especially for Cln2.

Measured band intensities vary between synchronization methods and, in turn, calculated absolute protein numbers vary as well. Band intensities and protein numbers are larger in hydroxyurea and nocodazole synchronized cells. Contrary to smFISH data where *CLB5* shows less than half of the mRNA numbers of *SIC1* (see Section 8.1), Sic1 and Clb5 reaches similar maximal protein levels in elutriated cells.

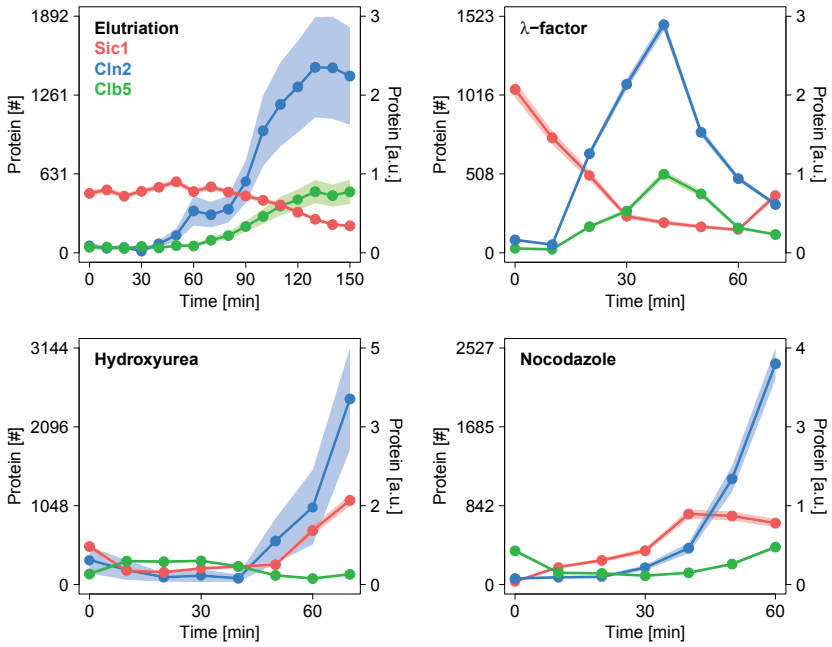

Figure 8.3: Pre-processed and normalized Western blot data of protein species Sic1, Cln2 and Clb5. In this figure, we represent protein time courses of the first cell cycle passage for elutriated, α-factor synchronized, hydroxyurea synchronized and nocodazole synchronized cell populations. Each panel shows mean values (dots) and errors (shaded area) which are calculated with the R function `alignME()` of the R package **blotIt2** before (right axis) and after (left axis) normalizing with absolute protein numbers given by the PaxDb database. Proteins are colored as follows: Sic1 in red, Cln2 in blue and Clb5 in green. The sampling rate is ten minutes. Lines between data points are for visualization only.

9. Mathematical modeling

9.1 Synchronization influences cell cycle timing and gene transcription

We estimated mRNA parameters from smFISH data in the mRNA optimization step and re-estimated these parameters from Western blot data in the protein optimization step (see Section 5.2). In Figure 9.1, we show results of 2000 stochastic simulations (see Section 4.4 for calculation details) for each mRNA species and synchronization specific parameter sets. Parameter estimation by using smFISH data was successful and predicted mRNA distributions reproduce Western blot data on the protein level.

Simulated mRNA distributions for parameters estimated from smFISH data ("No synchronization", dark gray) follow the behavior of smFISH data ("Data", colored). Whereas some simulated distributions almost perfectly reproduce smFISH data, others do not reproduce mean (white squares) or median (black lines in boxes). Simulated distributions of *CLN2* in early G1 phase, late G1 phase and P/M phase show the same tails and distributional widths as smFISH data. Similarly, simulated distributions of *CLB5* and smFISH data are almost identical from early G1 to G2 phase. Distributions of *SIC1* and *CLN2* in T/C phase are less well reproduced. mRNA distributions over the whole cell cycle are in good agreement with smFISH data (see Figure E.7).

Differences between simulated and measured distributions are not necessarily due to a bad parameter estimation. In Section 4.5, we illustrated artifacts resulting from simulating low molecular numbers and cell cycle phases enclosing either regions of in- or decreasing mRNA numbers or transitions between low and high transcription regions. Here, a χ^2 goodness of fit test is not appropriate. This test is highly sensitive and always lead to significant differences between measured and simulated data. Performing another 2000 stochastic simulations may reproduce smFISH data better or worse but would neither lead to a perfect fit.

Simulated distributions for re-estimated mRNA parameters from Western blot data which are measured for different synchronization methods ("Elutriation", "α-factor", "Hydroxyurea" and "Nocodazole", different grayscales) indicate variations in cell cycle timing and gene transcription. Hydroxyurea and nocodazole synchronized cells produce more mRNA molecules than predicted for unsynchronized cells or other synchronization methods. Larger mRNA numbers results from larger protein numbers shown in Figure 8.3 (see Section 8.3). Since we did not measure absolute protein numbers for used synchronization methods, we cannot exclude that differences in protein numbers are due to separate measurements. Normalizing Western blot data (see Section 3.4) changes scales from relative to absolute protein numbers but do not change relations between protein numbers of different synchronization methods.

In α-factor synchronized cells, *CLN2* shows a larger high transcription region which spans approximately two cell cycle phases (see Figure E.8 as well). In elutriated and hydroxyurea synchronized cells, the positioning and the length of high transcription regions are closest to unsynchronized cells. Since we measured cell cycle phase lengths for unsynchronized cells only, we cannot determine synchronization specific cell cycle phase lengths. As a simplification, we equaly scaled phase lengths. However, we predict how gene transcription has to be to reproduce protein numbers of synchronized cells.

There are simulated mRNA distributions for different synchronization methods which are similar to smFISH data meaning that estimated mRNA parameters are similar between unsynchronized and synchronized cells. Regarding the use of mRNA priors, similarities are to be expected, especially for parameters whose mRNA priors are close and allow for small deviations.

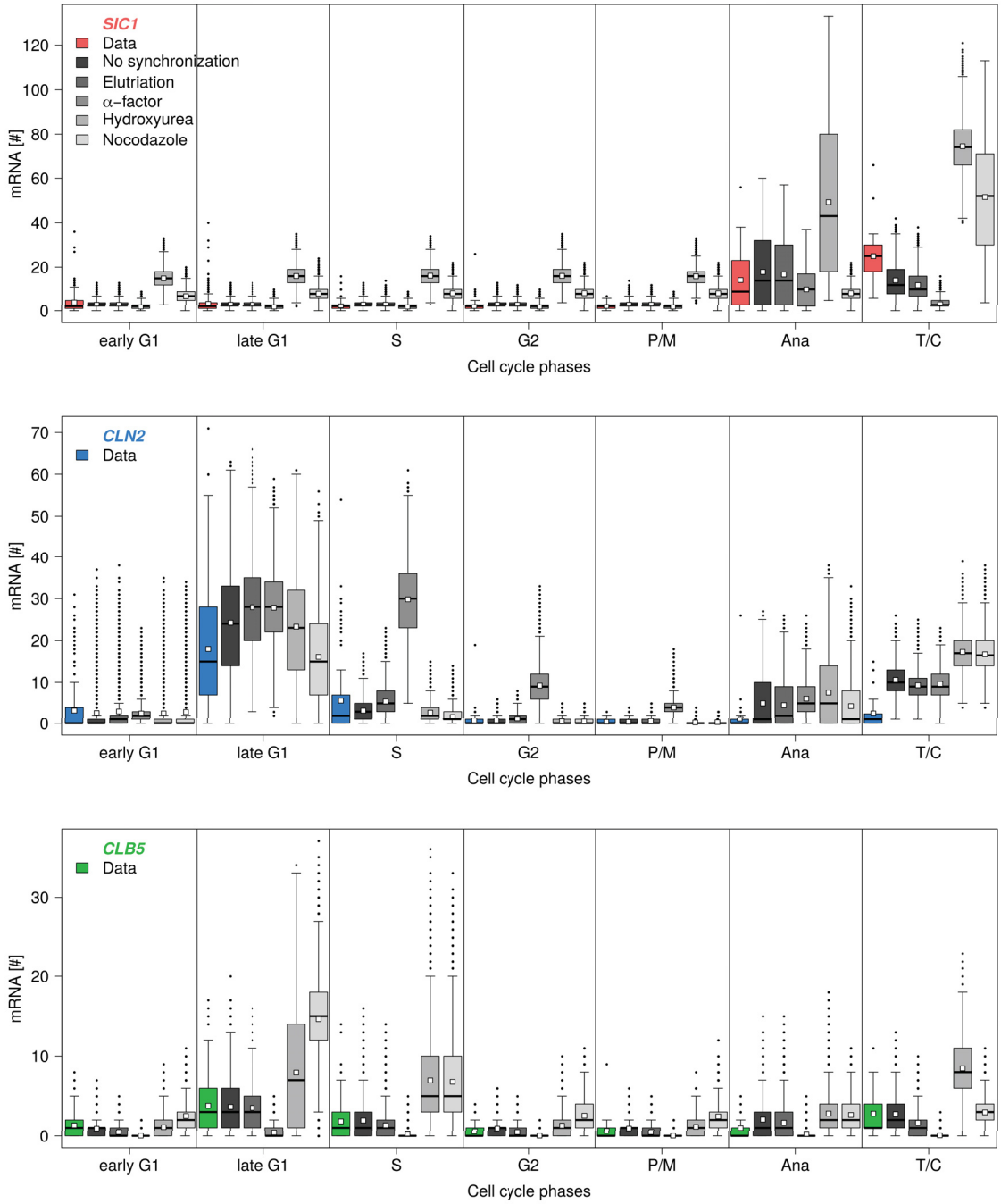

Figure 9.1: Measured and simulated phase-resolved mRNA distributions. In this figure, we show results of 2000 stochastic simulations (see Section 4.4 for calculation details) for mRNA species *SIC1* (red), *CLN2* (blue) and *CLB5* (green) and synchronization specific parameter sets (grayscales), named "No synchronization", "Elutriation", "α-factor", "Hydroxyurea" and "Nocodazole". smFISH data, named "Data", include measurements of more than 900 cells per mRNA species. We used integer valued phase lengths in the simulation (see Table 8.1) and simulated 129 time points ($\Delta t = 1$). "No synchronization" corresponds to simulations with parameters estimated from smFISH data in the mRNA optimization step. Other parameters are re-estimated from Western blot data in the protein optimization step. Distributions are represented as boxplots. First ($Q1$) and third quartile ($Q3$) form boxes with median (black lines) and mean (white squares). Whiskers range from $Q1 - 1.5 \cdot IQR$ to $Q3 + 1.5 \cdot IQR$ with an interquartile range of $IQR = Q3 - Q1$. Outliers are marked by black dots. We distinguished between seven phases: early G1, late G1, S, G2, pro-/metaphase (P/M), anaphase (Ana) and telophase/cytokinesis (T/C).

9.2 Observables successfully reproduce protein time courses of different synchronization methods

We estimated protein parameters from Western blot data which are measured for different synchronization methods in the protein optimization step (see Section 5.2). In Figure 9.2 we show how observables $\boldsymbol{Y}(t, \theta)$ (gray solid lines) fit to Western blot data (colored dots and error bars). Therefore, we numerically solved the ODE system defined in Equation 5.1 and calculated observables $Y_2(t, \theta)$, $Y_4(t, \theta)$ and $Y_6(t, \theta)$ according to observation functions given in Table 5.1.

We can reproduce protein time courses for differently synchronized cell populations. Sic1 time courses are best reproduced over all synchronization methods (first column). Sic1 time courses of hydroxyurea and nocodazole synchronized cells catch data points perfectly within errors. The only time course not covering maximal protein level shown in the data is Cln2 simulated for elutriated cells (first row). L_1 regularization which is discussed in Chapter 12 even worsens the data fit. The simulated protein time course lies outside error bars.

Protein time courses of nocodazole synchronized cells almost perfectly fit data points for every protein species (last row). However, Clb5 shows a down regulation between zero and ten minutes which disappears after applying L_1 regularization. A similar behavior occurs for time courses of hydroxyurea synchronized cells where Clb5 is up regulated in the same time window but survives L_1 regularization (Third row). These up and down regulations can be a result of using a changed cell cycle timing (see Figure E.3) which is dependent on the synchronization method.

Different synchronization methods require different initial settings to successfully fit Western blot data. Elutriated and α-factor synchronized cells arrest in G1 phase with a maximal Sic1 level which predominantly decrease over the cell cycle. We decided to start ODE simulation in anaphase where Sic1 is high according to our smFISH data. The first data point still arises in G1 phase. Additionally, we did not estimate initial values. It was not possible to determine a Sic1 production rate if initial values are estimated.

Western blot data of hydroxyurea and nocodazole synchronized cells need a different setup. ODE simulation and data start in S and G2 phase, respectively. We estimated initial values for two reasons. First, Sic1 levels emerge from a single high transcription region on the mRNA level but there are two high Sic1 levels in hydroxyurea synchronized cells. An initial value fixed to zero generates a second peak. Second, protein time courses of Cln2 (hydroxyurea) and Clb5 (nocodazole) are similar to Sic1 (hydroxyurea) but with a second high transcription region on the mRNA level. Estimating initial values prevents an additional peak as well.

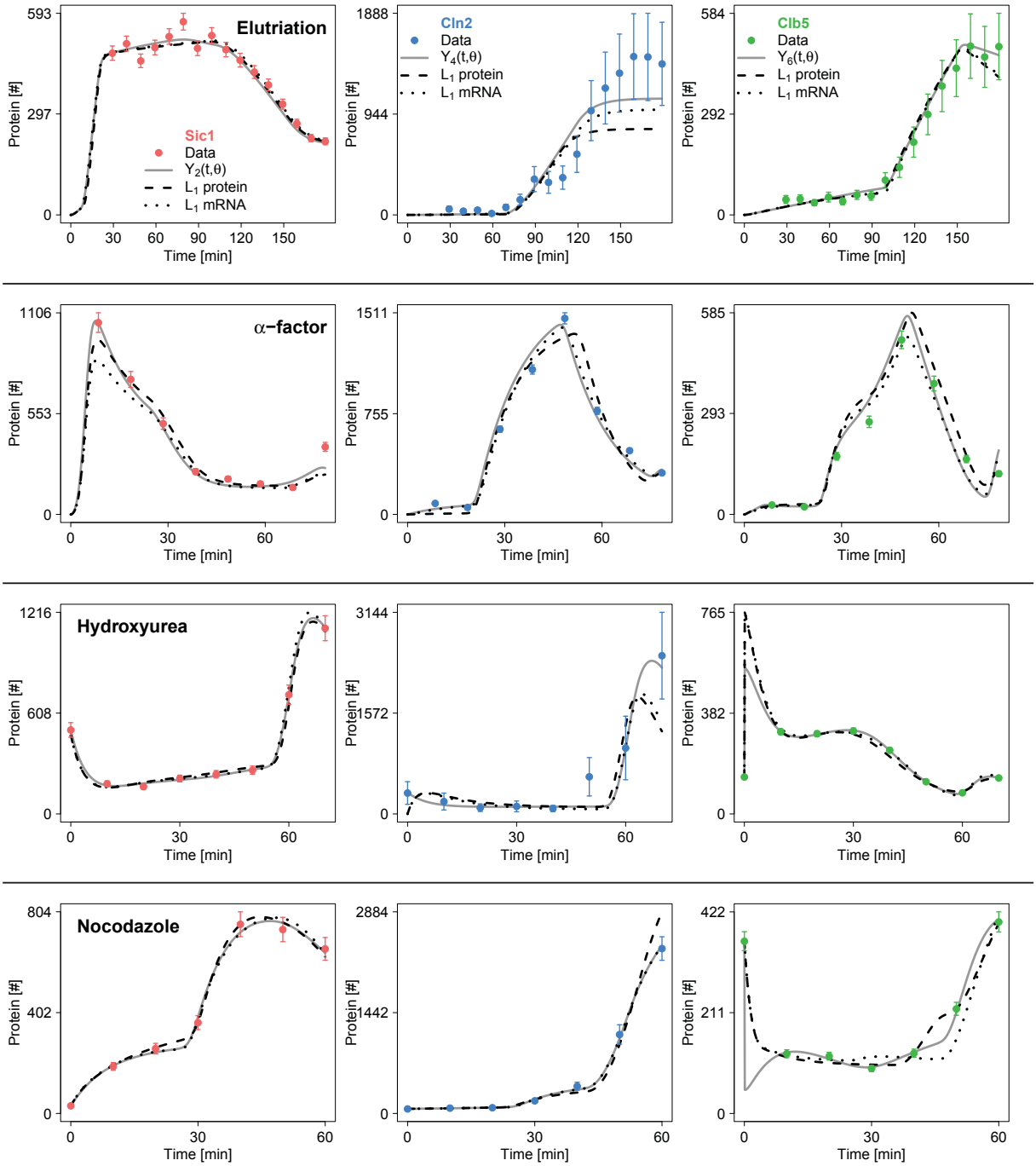

Figure 9.2: Measured and simulated protein time courses . In this figure, we illustrate observables $Y_2(t,\theta)$, $Y_4(t,\theta)$ and $Y_6(t,\theta)$ (gray solid lines, see Table 5.1) fitted to Western blot data (colored dots and error bars) which are measured for different synchronization methods (elutriation in the first row, α-factor in the second row, hydroxyurea in the third row and nocodazole in the fourth row). Sic1 (red) is given in the first, Cln2 (blue) in the second and Clb5 (green) in the third column. Plotted data points and error bars are results from data pre-processing (see Section 3.3) and normalization (see Section 3.4) as shown in Figure 8.3. Fitting results after applying L_1 regularization to protein parameters (black dashed lines) and mRNA fold changes (black dotted lines) are marked as well.

10. Parameter estimation

10.1 mRNA optimization perform superior to protein optimization

We estimated mRNA and protein parameters in a 2-step-optimization by using a multi-start local optimization (see Section 5.4). mRNA parameters are estimated from smFISH data in the mRNA optimization step and re-estimated from Western blot data in the protein optimization step. Protein parameters are estimated from Western blot data in the protein optimization step. In Figure 10.1, we present the performance of the mRNA optimization step dependent on the mRNA species and of the protein optimization step dependent on the synchronization method. We initialized 1000 or 2000 optimization runs and sorted values of the objective function by size. In case of optimizing the negative log-likelihood function, smallest values refer to the optimum. Table 10.1 includes numbers of successful optimizations.

In the mRNA optimization step, *SIC1* optimization shows the smallest number of successful optimization runs but almost every successful optimization leads to the same value of the objective function which indicates a good performance and the closeness to the global optimum. About 800 initializations successfully finished in *CLN2* optimization. Compared to *SIC1*, several local optima arise. *CLB5* optimization shows the worst performance but nearly every optimization run was successful. Even if there is a clear minimum, values of the objective function increase almost continuously for the most part.

In the protein optimization step, the overall performance is inferior to the mRNA optimization step. Less than a third of initialized optimization runs successfully finished. We got several local minima which are hard to distinguish. Values of the objective function are different and separated in some regions but also continuously increasing in others. The optimization related to α-factor synchronization is the only optimization where the smallest value is actually reached more than once. For other synchronization methods, only the second-smallest value is obtained several times. We started 2000 optimization runs for optimizations related to hydroxyurea and nocodazole synchronization to improve performance.

Optimization	*SIC1*	*CLN2*	*CLB5*	Elutriation	α-factor	Hydroxyurea	Nocodazole
Successes	374	798	961	272 (307)	323 (420)	483 (961)	653 (929)

Table 10.1: Number of successful optimization runs. This table shows the number of successful optimization runs out of 1000 (*SIC1*, *CLN2*, *CLB5*, elutriation and α-factor) or 2000 (hydroxyurea and nocodazole) initializations. mRNA parameters are individually estimated for each mRNA species in the mRNA optimization step. mRNA and protein parameters are estimated all at once for each synchronization method in the protein optimization step. Numbers in brackets mark optimizations that did not converge.

10.2 Parameters related to cell cycle timing are most clearly affected by synchronization

In Figure 10.2, we plot mRNA parameters of unsynchronized cells estimated from smFISH data in the mRNA optimization step ("No") and mRNA parameters of synchronized cells re-estimated from Western blot data in the protein optimization step ("Elut.", "α", "Hydro." and "Noco."). Error bars mark 95% confidence intervals which are calculated from profile likelihoods. Profile likelihoods are discussed in Chapter 11. Caused by its closed confidence interval, we present parameter values of the scaling factor introduced to combine mRNA and protein optimization and estimated from Western blot data in the protein optimization step as well (see Section 5.3). Parameter values are given in Table F.2.

Synchronization specific mRNA parameters are uniquely determined and show clear variations in start and end times of high transcription regions. In contrast, most mRNA reaction rates are

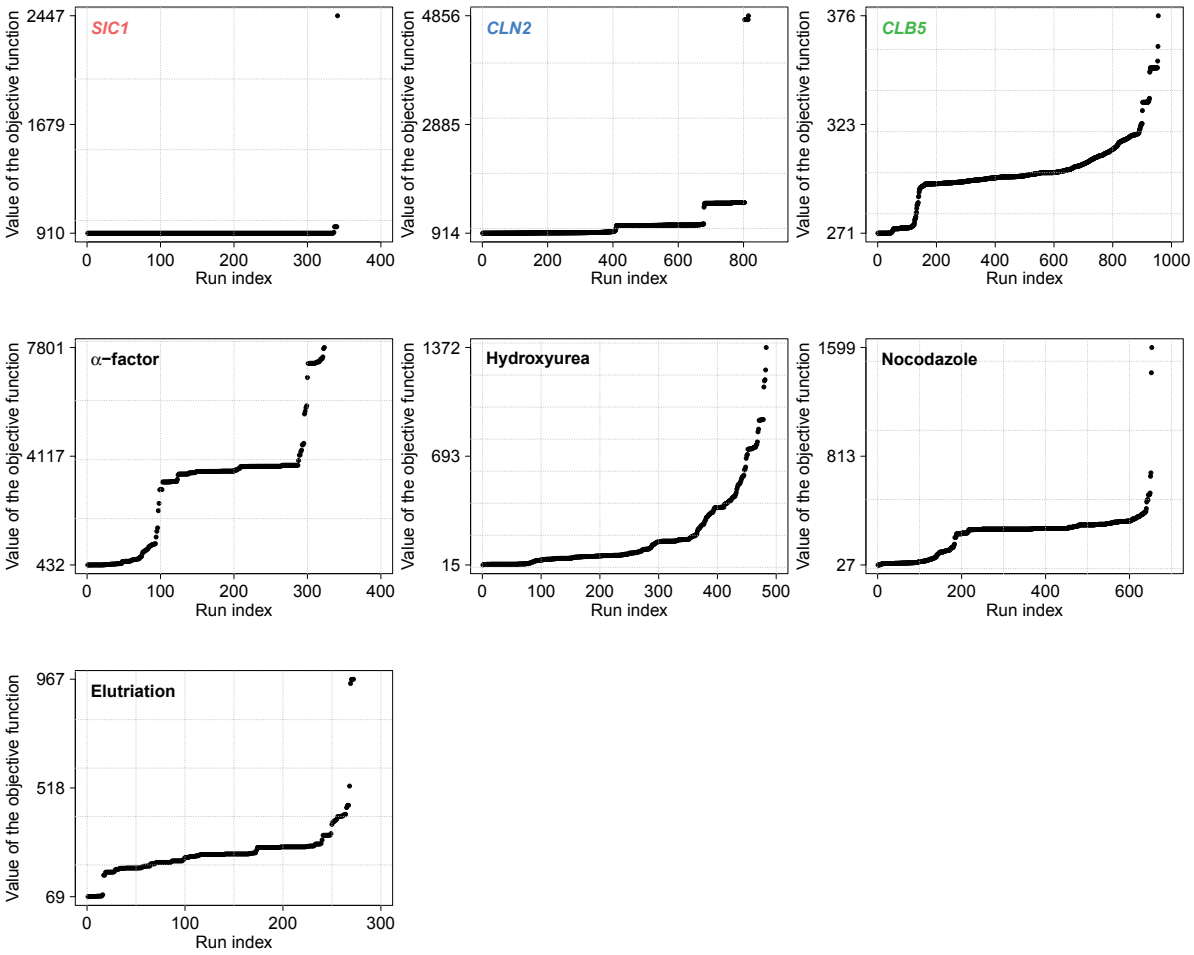

Figure 10.1: Performance of the 2-step-optimization. In this figure, we show the performance of the multi-start local optimization in the mRNA (first row) and the protein optimization step (second and third row). mRNA parameters are individually estimated for each mRNA species in the mRNA optimization step. mRNA and protein parameters are estimated all at once for each synchronization method in the protein optimization step. We started 1000 (*SIC1*, *CLN2*, *CLB5* elutriation and α-factor) or 2000 (hydroxyurea and nocodazole) optimization runs. In these likelihood waterfall plots, values of the respective objective functions are sorted by size. We included converged optimization runs only (see Table 10.1).

less clearly distinguishable due to its broader confidence intervals. In Section 9.1, we have seen synchronization specific variations in gene transcription which seems to be a result of differences in timing. Three parameters ($t_{s_{1,0}}$, $t_{s_{3,e,2nd}}$ and $k_{1,high}$) estimated for unsynchronized cells have an one-sided open confidence interval which closes for smaller confidence levels (see Figure E.9). Non-overlapping confidence intervals indicate significant differences between parameter values. If confidence intervals overlap, parameter values can be significantly or not significantly different [117]. Unfortunately, there is no appropriate statistical test to judge significance.

In Figure 10.3, we show protein parameters of synchronized cells estimated from Western blot data in the protein optimization step before (black) and after (gray) applying L_1 regularization (see Tables F.3 and F.4 as well). As discussed in Section 9.2, initial values are estimated for hydroxyurea and nocodazole synchronized cells only. Parameters k_7 and $C_{1,0}$ are not missing for α-factor and hydroxyurea synchronization but their values are atypically large compared to remaining parameter values and, therefore, are outside the plotting region. Unlike confidence intervals of mRNA parameters, dashed lines indicate confidence borders going to infinity due to the Gaussian prior (see Section 6.2).

Synchronization specific protein parameters are not uniquely determined in most cases and are highly variable. It is not possible to assess significant differences between parameter values. A small number of confidence intervals closes for smaller confidence levels (see Figure E.10). In Chapter 12, we discuss how L_1 regularization applied to protein parameters will considerably improve parameter estimates (see Figure E.11 as well).

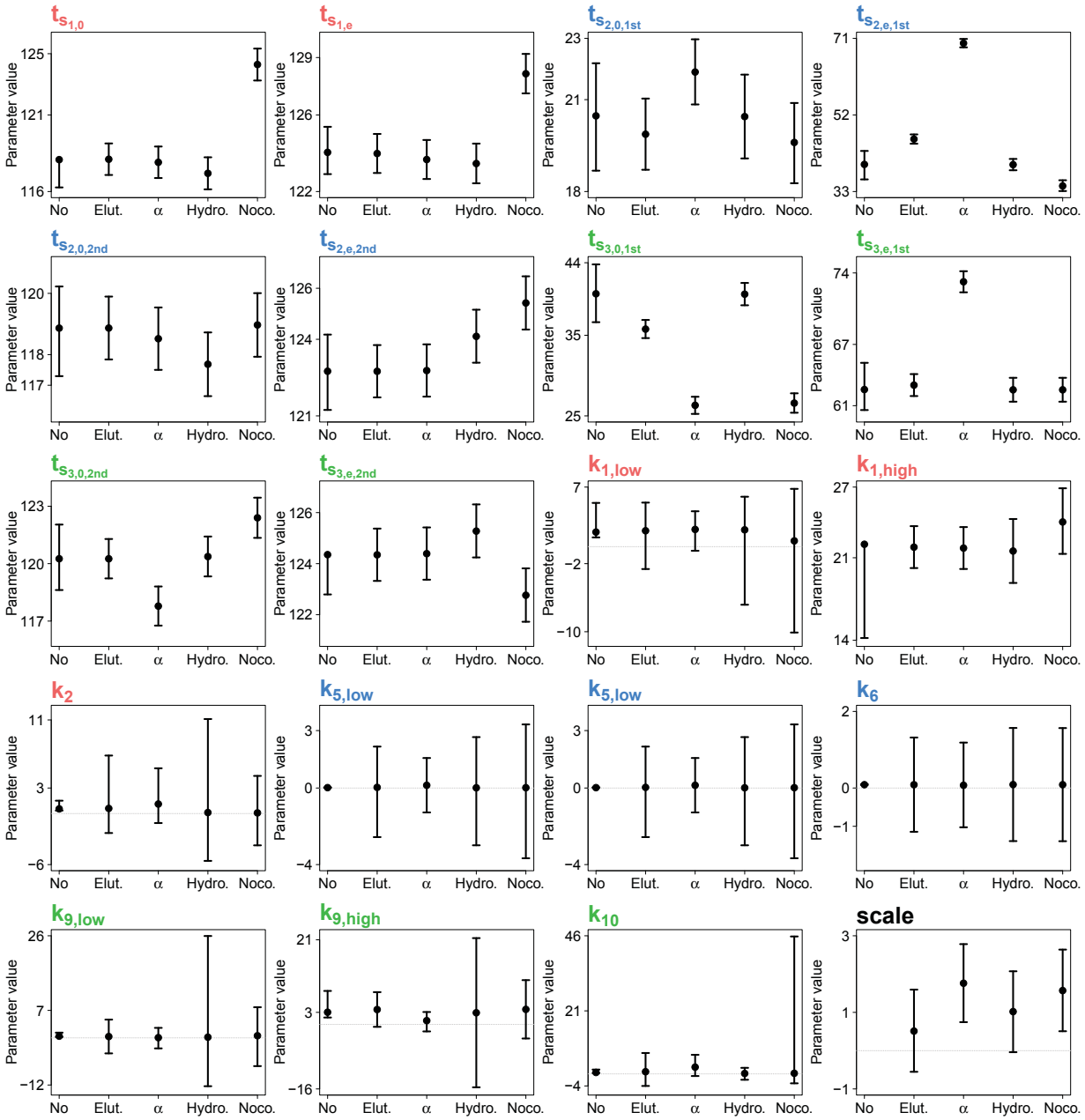

Figure 10.2: Estimated mRNA parameters. In this figure, we show estimated mRNA parameters and the scaling factor introduced to combine mRNA and protein optimization (see Tables 4.2 and 4.1 for associated reactions). mRNA parameters are estimated from smFISH data in the mRNA optimization step and marked as "No" (abbreviation for "No synchronization"). mRNA parameters are re-estimated from Western blot data in the protein optimization step and assigned to respective synchronization methods: "Elut." (abbreviation for "Elutriation"), "α" (abbreviation for "α-factor"), "Hydro." (abbreviation for "Hydroxyurea") and "Noco." (abbreviation for "Nocodazole"). The scaling factor is estimated in the protein optimization step only and , therefore, has no estimate for "No". Error bars represent 95% confidence intervals calculated from profile likelihoods. If confidence intervals were not determinable to both sides, we plotted an one-sided open confidence interval. A gray dotted line indicates zero. Parameter values are given in Table F.2.

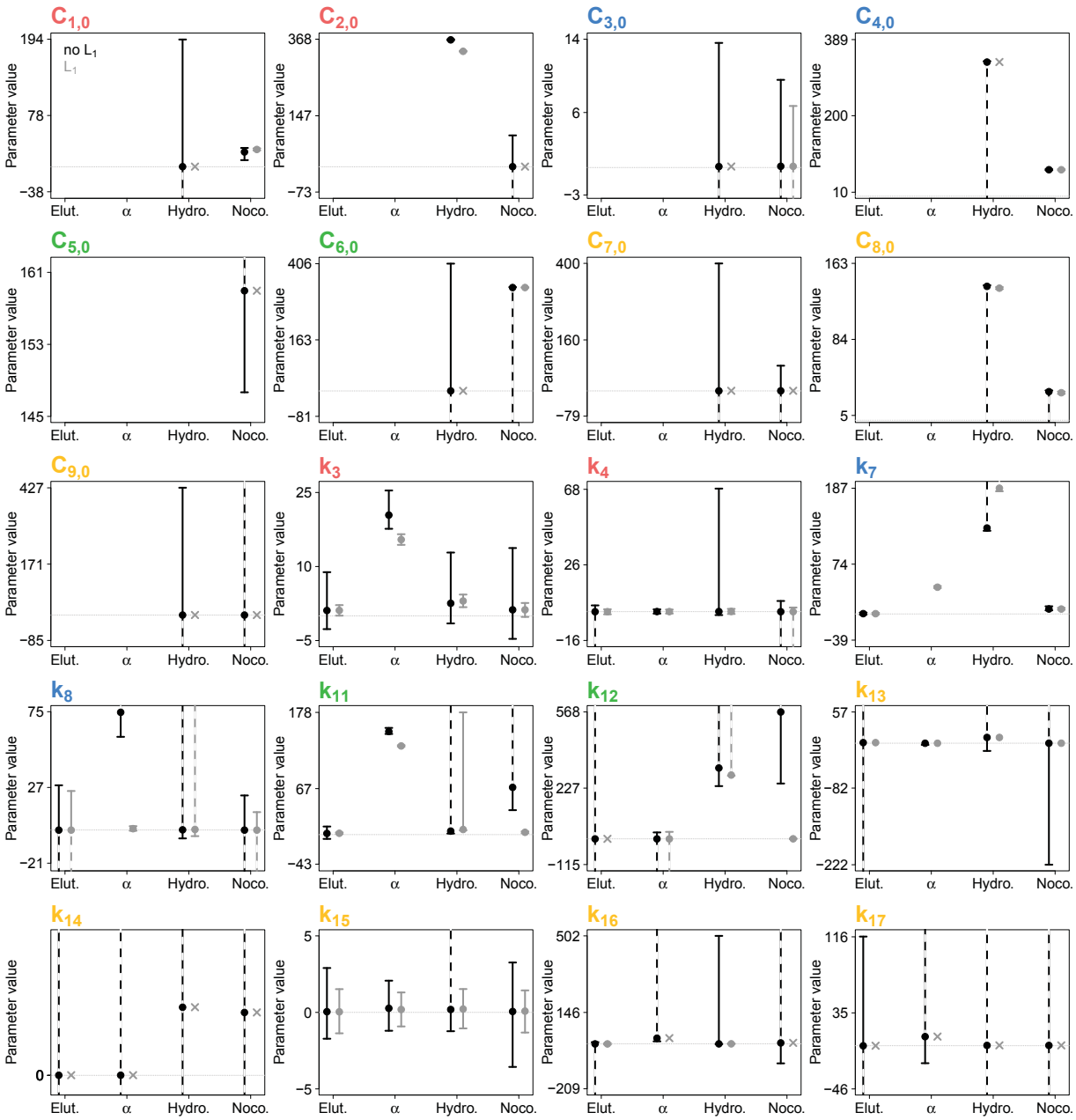

Figure 10.3: Estimated protein parameters. In this figure, we show protein parameters (see Tables 4.2 and 4.1 for associated reactions) estimated from Western blot data in the protein optimization step before (black) and after applying L_1 regularization (gray). Parameters which are removed during L_1 regularization show a cross instead of a dot with error bars. Parameter estimates are assigned to respective synchronization methods: "Elut." (abbreviation for "Elutriation"), "α" (abbreviation for "α-factor"), "Hydro." (abbreviation for "Hydroxyurea") and "Noco." (abbreviation for "Nocodazole"). Initial values are estimated for synchronization by hydroxyurea and nocodazole only. Parameters k_7 and $C_{5,0}$ are not missing for α-factor and hydroxyurea synchronization but their values are larger than the others. Error bars represent 95% confidence intervals calculated from profile likelihoods. Dashed lines indicate infinite confidence intervals. A gray dotted line indicates zero. Parameter values are given in Tables F.3 and F.4.

11. Identifiability analysis

11.1 mRNA parameters are generally identifiable

We calculated profile likelihoods for mRNA parameters estimated from smFISH data in the mRNA optimization step (see Equation 6.5) and re-estimated from Western blot data in the protein optimization step (see Equation 6.7) to judge identifiability and to determine confidence intervals (see Section 6.2). Confidence intervals were already presented in the previous section (see Section 10.2). Independent of the optimization step, mRNA parameters are generally identifiable. In Figure 11.1, we show illustrative examples of profile likelihoods of the mRNA optimization step. A comprehensive representation of all profile likelihoods is given in Figure E.12. Profile likelihoods of the mRNA optimization step have a single contribution referring to smFISH data.

Parameters related to mRNA species *CLN2* (blue) have almost perfectly quadratic profile likelihoods (example: $t_{s_{2,e,2nd}}$). For this reason, confidence intervals are symmetric and become equidistantly smaller for smaller confidence levels (see Figure E.9). These parameters are always identifiable. Parameters related to mRNA species *SIC1* (red) and *CLB5* (green) show in some cases a non-perfect quadratic profile likelihood which still crosses the 95% confidence level (example: $k_{9,high}$). Therefore, parameters are identifiable but confidence intervals are asymmetric.

Parameter $t_{s_{3,e,2nd}}$ is the only practically non-identifiable parameter regarding confidence levels larger than 90%. This parameter is still identifiable to a certain degree. Profile likelihoods of *SIC1* parameters are partly noisy indicating convergence problems (example: $t_{s_{1,0}}$). As parameter $k_{1,high}$ (see Figure E.12), profile likelihoods do not always cross confidence levels larger than 68%. Still, considering a larger parameter range leads to a broad asymmetric confidence interval for larger confidence levels. Even if parameters are identifiable, broad confidence intervals are not preferable.

In Figure 11.2, we show illustrative examples of profile likelihoods of the protein optimization step for α-factor synchronization. A comprehensive representation of all profile likelihoods is given in Figure E.14. Profile likelihoods calculated for elutriation, hydroxyurea synchronization and nocodazole synchronization are presented in Figures E.13, E.15 and E.16. Profile likelihoods of the protein optimization step have three contributions to the total profile. The data contribution refers to Western blot data. Additionally, the Gaussian and mRNA prior contributions feed in the total profile.

We use total profiles to evaluate identifiabilities and to calculate confidence intervals. Re-estimated mRNA parameters have predominantly quadratic profiles and, therefore, symmetric 95% confidence intervals. The Gaussian prior only contributes to parameters k_{10} and $k_{9,high}$ for hydroxyurea synchronization and parameter k_{10} for nocodazole synchronization. In the next section, we discuss how re-estimated mRNA parameters are influenced by Western blot data.

11.2 Cell cycle timing and gene transcription are equally affected by Western blot data

We determined contributions of Western blot data to the total profile of re-estimated mRNA parameters by using classifications introduced in Section 6.3. In Figure 11.2, we show four different data contributions. A total profile which is determined by the mRNA prior indicates shared parameter values among unsynchronized and synchronized cells (example: $t_{s_{1,0}}$). A data contribution, in turn, indicates a mismatch between them. Data can either contribute to one (examples: $t_{s_{2,e,1st}}$, $k_{9,high}$) or to two sides (example: $t_{s_{2,0,1st}}$) of the parameter re-estimate and can be strong (example: $t_{s_{2,e,1st}}$) or weak (examples: $t_{s_{2,0,1st}}$, $k_{9,high}$).

Even though we have seen in Section 10.2 that timing parameters differentiate more clearly compared to reaction rates, both are afftected to the same amount by Western blot data. Confi-

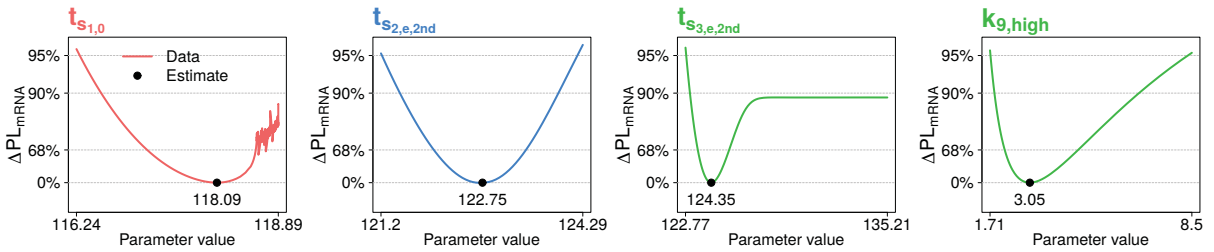

Figure 11.1: Selection of profile likelihoods representing mRNA parameters estimated in the mRNA optimization step. In this figure, we show illustrative examples of profile likelihoods of mRNA parameters (see Tables 4.2 and 4.1 for associated reactions) estimated from smFISH data in the mRNA optimization step as shown in Figure 6.1. A comprehensive representation of all profile likelihoods is given in Figure E.12. Smallest, largest and estimated parameter values are plotted on the x-axis. Likelihood ratios and different confidence levels are given on the y-axis. Colored lines show contributions of the data. Red, blue and green refer to *SIC1*, *CLN2* and *CLB5*, respectively. Parameter estimates are marked as black dots.

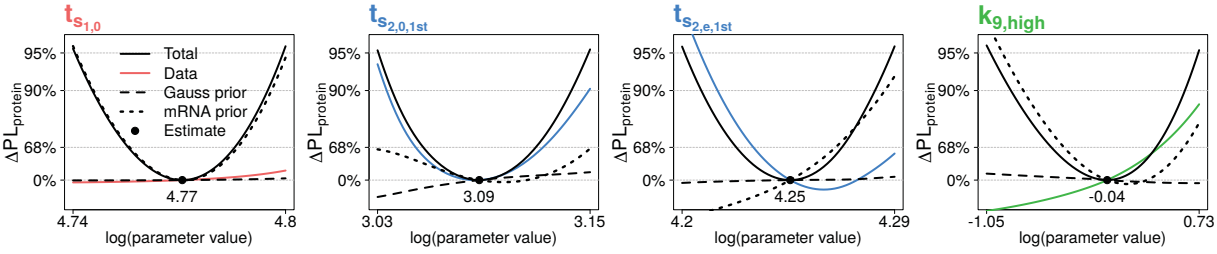

Figure 11.2: Selection of profile likelihoods representing mRNA parameters re-estimated in the protein optimization step for α-factor synchronized cells. In this figure, we show illustrative examples of profile likelihoods of mRNA parameters (see Tables 4.2 and 4.1 for associated reactions) re-estimated from Western blot data in the protein optimization step as shown in Figure 6.3. A comprehensive representation of all profile likelihoods is given in Figure E.14. Smallest, largest and estimated log-transformed parameter values are plotted on the x-axis. Likelihood ratios and different confidence levels are given on the y-axis. Colored lines show contributions of the data. Red, blue and green refer to Sic1, Cln2 and Clb5, respectively. Contributions of the Gaussian and mRNA prior are illustrated by dashed and dotted black lines. The total profile is given by a black solid line. Parameter estimates are marked as black dots.

dence intervals of timing parameters become smaller (closer profile likelihoods) if data contributions are present. Table 11.1 summarize synchronization specific data contributions to total profiles. Strong effects are marked by arrowheads and colors indicate LHS (orange), RHS (light blue), LRHS (gray) and no (white) effects.

The optimum of the sum of quadratic contributions of mRNA prior and data, which are shifted against each other, lies between optima of individual contributions. A data contribution shifted to the left leads to a smaller optimum compared to the optimum of the mRNA prior and vice versa. This is the reason why data effects on the LHS mostly lead to larger parameter values whereas effects on the RHS lead to smaller parameter values compared to those of the mRNA optimization step. As long as we have no statistical evidence for parameter differences, we cannot statistically verify this trend.

Parameter	Synchronization			
	Elutriation	α-factor	Hydroxyurea	Nocodazole
$t_{s_{1,0}}$				
$t_{s_{1,e}}$				
$t_{s_{2,0,1st}}$				
$t_{s_{2,e,1st}}$	$<$	$<$		
$t_{s_{2,0,2nd}}$				
$t_{s_{2,e,2nd}}$				
$t_{s_{3,0,1st}}$		$>$		
$t_{s_{3,e,1st}}$		$<$		
$t_{s_{3,0,2nd}}$				
$t_{s_{3,e,2nd}}$				
$k_{1,low}$				
$k_{1,high}$				
k_2			$>$	
$k_{5,low}$	$<$			
$k_{5,high}$				
k_6		$<$		
$k_{9,low}$				
$k_{9,high}$				
k_{10}				
$scale$				$<$

Non LHS RHS LRHS

Table 11.1: Data effects to mRNA profile likelihoods of the protein optimization step. This table shows effects of Western blot data to mRNA parameters re-estimated in the protein optimization step for elutriation, α-factor synchronization, hydroxyurea synchronization and nocodazole synchronization. For reasons of consistency, we show data contributions to the scaling factor as well. An effect on the LHS is colored in orange, on the RHS in light blue and on both sides (LRHS) in gray. A white cell indicates that data do not contribute the total profile. Arrowheads mark strong effects either on the LHS ($<$) or on the RHS ($>$). A detailed description of possible data contributions is given in Figure 6.4.

11.3 Protein parameters are predominantly practically non-identifiable

We calculated profile likelihoods for protein parameters estimated from Western blot data in the protein optimization step (see Equation 6.7) to judge identifiability and to determine confidence intervals which are presented in Section 10.2 (see Section 6.2). Independent of the synchronization method, the majority of protein parameters is non-identifiable and practical non-identifiabilities predominate.

In Figure 11.3, we show illustrative examples of profile likelihoods of the protein optimization

step for synchronization by elutriation. A comprehensive representation of all profile likelihoods is given in Figure E.17. Profile likelihoods calculated for elutriation, hydroxyurea synchronization and nocodazole synchronization are presented in Figures E.18, E.19 and E.20.

A small number of parameters is identifiable (examples: k_{11} and k_{15}) meaning that the contribution of the Gaussian prior is negligible. For an even smaller number, the total profile is determined by the data (example: k_{15}) and not by the mRNA prior (example: k_{11}). Practically non-identifiable parameters are characterized by a total profile determined to one side by the Gaussian prior (example: k_8). In the worst case, parameters are structurally non-identifiable (example: k_{12}). In case of a practical non-identifiability and a total profile governed by the mRNA prior, data are insufficient to determine parameter values.

Table 11.2 summarizes identifiabilities of protein parameters. Non-quadratic shapes of total profiles are marked with a asterisks in the column of the 95% confidence level and colors highlight identifiable (white), practically non-identifiable (light blue) and structurally non-identifiable parameters. Not estimated initial values for elutriated and α-factor synchronized cells are colored in light gray.

Practically non-identifiable parameters dominates for almost every confidence level. In some cases, identifiability is dependent on the confidence level so that smaller confidence levels change structurally to practically non-identifiable parameters and practically non-identifiable to identifiable parameters (see Figure E.10 as well). There are parameters with the same identifiablitiy over all synchronizations (example: k_3) and with different identifiabilities (example: k_{11}). Deviations from quadratic shaped total profiles can be a result of an optimum not necessarily representing the global optimum.

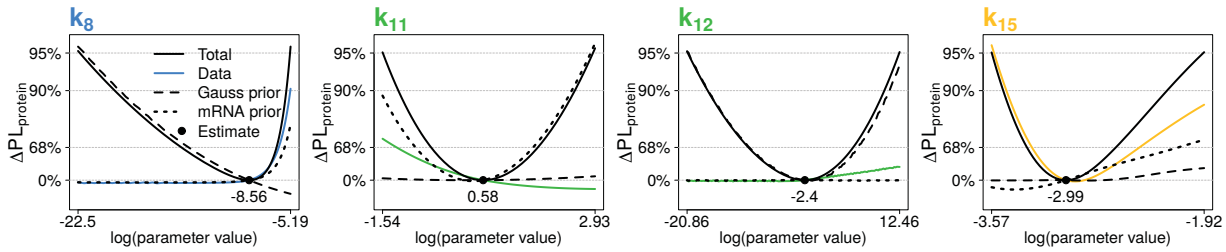

Figure 11.3: Selection of profile likelihoods representing protein parameters estimated in the protein optimization step for elutriated cells. In this figure, we show illustrative examples of profile likelihoods of protein parameters (see Tables 4.2 and 4.1 for associated reactions) estimated from Western blot data in the protein optimization step as shown in Figure 6.3. A comprehensive representation of all profile likelihoods is given in Figure E.17. Smallest, largest and estimated log-transformed parameter values are plotted on the x-axis. Likelihood ratios and different confidence levels are given on the y-axis. Colored lines show contributions of the data. Blue and green refer to Cln2 and Clb5, respectively. Yellow represents a mixture of Sic1 and Clb5. Contributions of the Gaussian and mRNA prior are illustrated with dashed and dotted black lines. The total profile is given by a black solid line. Parameter estimates are marked as black dots.

11.4 mRNA priors contribute less to protein parameters of α-factor synchronized cells

We determined the contribution of the mRNA prior to the total profile of protein parameters by using classifications introduced in Section 6.3. In Figure 11.3, we show four different mRNA contributions. The mRNA prior can contribute to both sides of the parameter estimate (example: k_{11}) or to one side (examples: k_8 and k_{15}) and can be strong (example: k_{11}) or weak (examples: k_8 and k_{15}). A parameter is determined by the Gaussian prior and/or the data if the mRNA prior does not contribute (example: k_{12}). A contribution of the mRNA prior not necessarily excluded a data contribution (example: k_8).

A few parameters have the same mRNA prior contribution over all synchronizations (example: k_{17}). Parameter k_3 is the only protein parameter determined as being identifiable over all synchronizations and confidence levels (see Table 11.2) and at the same time has a mRNA prior

Par	Synchronization											
	Elutriation			α-factor			Hydroxyurea			Nocodazole		
	95%	90%	68%	95%	90%	68%	95%	90%	68%	95%	90%	68%
$C_{1,0}$												
$C_{2,0}$												
$C_{3,0}$												
$C_{4,0}$												
$C_{5,0}$										*		
$C_{6,0}$										*		
$C_{7,0}$												
$C_{8,0}$							*			*		
$C_{9,0}$												
k_3												
k_4												
k_7												
k_8												
k_{11}												
k_{12}												
k_{13}												
k_{14}	*									*		
k_{15}							*					
k_{16}							*					
k_{17}												

Not Estimated | Identifiable | Structural | Practical

Table 11.2: Protein parameter identifiabilities. This table shows identifiabilities for different confidence levels which are determined from total profiles of protein parameters estimated in the protein optimization step for elutriation, α-factor synchronization, hydroxyurea synchronization and nocodazole synchronization (see Figures E.17, E.18, E.19 and E.20). Structurally non-identifiable parameters are colored in orange and practically non-identifiable parameters are colored in light blue. Identifiable parameters are given in white. Initial values colored in light gray are not estimated. Asterisks in the 95% column mark total profiles deviating from a quadratic shape.

contribution mostly to both sides. Table 11.3 summarizes synchronization specific mRNA prior contributions to total profiles. Strong effects are marked by arrowheads and colors indicate LHS (orange), RHS (light blue), LRHS (gray) and no (white) effects. Not estimated initial values for elutriated and α-factor synchronized cells are colored in light gray.

Protein parameters estimated for α-factor synchronized cells show less mRNA prior contributions compared to the other synchronizations. As seen in the previous section, this fact has no effect on parameter identifiability but suggest a stronger data contribution.

Parameter	Synchronization			
	Elutriation	α-factor	Hydroxyurea	Nocodazole
$C_{1,0}$				$<$
$C_{2,0}$				
$C_{3,0}$				
$C_{4,0}$				
$C_{5,0}$				
$C_{6,0}$				
$C_{7,0}$				
$C_{8,0}$				
$C_{9,0}$				
k_3		$>$		$>$
k_4				$>$
k_7				
k_8				
k_{11}	$>$	$>$		
k_{12}				
k_{13}				
k_{14}				
k_{15}				
k_{16}				
k_{17}				

Not Estimated | Non | LHS | RHS | LRHS

Table 11.3: mRNA prior effects to protein profile likelihoods of the protein optimization step. This table shows mRNA prior effects to protein parameters estimated in the protein optimization step for elutriation, α-factor synchronization, hydroxyurea synchronization and nocodazole synchronization. An effect on the LHS is colored in orange, on the RHS in light blue and on both sides (LRHS) in gray. A white cell indicates that the data have no effect on the total profile. Initial values colored in light gray are not estimated. Arrowheads mark strong effects either on the LHS ($<$) or on the RHS ($>$). A detailed description of possible data contributions in given in Figure 6.4.

11.5 Parameter dependencies persist between identifiable and non-identifiable parameters

We analyzed parameter dependencies by plotting parameter values θ_ι against parameter values of the profiled parameter $\theta_{\underline{\iota}}$ (see Section 6.1). Parameters are dependent if variances of parameter values θ_ι are larger or smaller than ± 0.01. In Figures 11.4, 11.5 and 11.6, we present examples of parameter dependencies for two identifiable parameters ($k_{5,high}$ and $k_{9,high}$) of the mRNA level as well as a structurally non-identifiable parameter (k_{14}) of the protein level. Parameters estimated in the mRNA optimization step are named "No synchronization". Parameters of the protein optimization step are named by its synchronization method ("Elutriation", "α-factor", "Hydroxyurea" and "Nocodazole").

Independent parameters show a flat line (gray) and dependent parameters an in- or decreasing

line (black). Changing the selecetion criterion can change the number of dependent parameters. Other reasonable criteria are variances larger or smaller than ± 0.05 or ± 0.1. Dependent on the selection criterion, parameters can erroneously be marked as being independent (see "Nocodazole" in Figure 11.4) or dependent. The latter occurs if we choose small variances as selection criterion. Different selection criteria, e.g. average percentage distance from the mean or estimated parameter value, give similar results. Some detected parameter dependencies do not have in- or decreasing lines and their curves have non-well defined shapes. Additionally, non-quadratic profile likelihoods cause worse curves.

Parameters along the profile likelihood of parameter $k_{5,high}$ estimated for unsynchronized cells behave as expected (see Figure 11.4). The parameter is identifiable and shows no parameter dependencies. The same parameter is identifiable for every synchronization method and parameter dependencies are still detected. Parameter dependencies mainly persist between parameters from the same symmetry group (see Table F.5). Nocodazole has an additional parameter dependency to parameter k_8 which is a practically non-identifiable parameter of the protein level. The reason why hydroxyurea shows no parameter dependencies is that the total profile is determined by the mRNA prior and the mRNA prior is close and symmetric.

mRNA parameter $k_{9,high}$ shows parameter dependencies in four synchronizations (see Figure 11.5). Parameter dependencies detected for unsynchronized cells represent asymmetric profile likelihoods. In contrast, parameter $k_{5,high}$ shows almost perfect quadratic profile likelihoods. Hydroxyurea shows again no parameter dependencies for the same reasons. Beside parameter dependencies expected by symmetry groups, elutriation and nocodazole show additional parameter dependencies to protein parameters k_{12} and k_{16} which are practically non-identifiable. Even if symmetry groups are the same among synchronization, parameter dependencies differ.

Parameter dependencies between protein parameters do not only occur between structurally non-identifiable parameters, do not detect every structurally non-identifiable parameter and are not necessarily the same for different profile likelihoods. In Figure 11.6, we show synchronization specific dependencies to parameter k_{14} related to the protein optimization step. For instance, elutriation miss structurally non-identifiable parameters k_{12} and k_{13} and additionally detected an identifiable parameter of the mRNA level ($k_{1,low}$). Parameters along the profile likelihood of parameter k_{14} show a dependency to parameter k_{17} for nocodazole but not the other way around (see Figure E.23). Parameters referring to initial values detect dependencies to initial values only.

11.6 Variations in model trajectories occur equally for identifiable and non-identifiable parameters

We analyzed variability of variables $\boldsymbol{C}(t, \theta)$ and observables $\boldsymbol{Y}(t, \theta)$ (see Tables 4.1 and 5.1) by plotting trajectories for parameter values along profile likelihoods (see Section 6.1). Trajectories reveal regions where parameter uncertainty has the largest influence on the chemical reaction system and which trajectories cannot be determined at all. In Figures 11.7, 11.8 and 11.9, we present examples of model trajectories related to α-factor synchronization for an identifiable parameter of the mRNA level ($k_{5,high}$), a practically (k_7) and a structurally non-identifiable (k_{14}) parameter of the protein level. Parameters are estimated from Western blot data in the protein optimization step.

Trajectories of observables show variation along the profile likelihood of the structurally non-identifiable parameter k_{14} and no variation along profile likelihoods of the identifiable parameter $k_{5,high}$ and practically non-identifiable parameter k_7. Nevertheless, variations occur for identifiable (see Figure E.24) or practically non-identifiable (see Figure E.25) parameters as well. These variations are rather small compared to variations arising in trajectories of variables.

Regarding trajectories of variables along the profile likelihood of the identifiable parameter $k_{5,high}$ (see Figure 11.7), variation in variable $C_3(t, \theta)$ directly reflects the profile likelihood which is determined by the mRNA prior and the data contribution is negligible (see Figure E.14). Variations in variables $C_5(t, \theta)$ and $C_9(t, \theta)$ are very small compared to variable $C_3(t, \theta)$. In variables $C_3(t, \theta)$ and $C_5(t, \theta)$, smaller values of parameter $k_{5,high}$ (light gray lines) also lead to smaller molecule numbers. Differently, larger parameter values of $k_{5,high}$ (gray lines) lead to

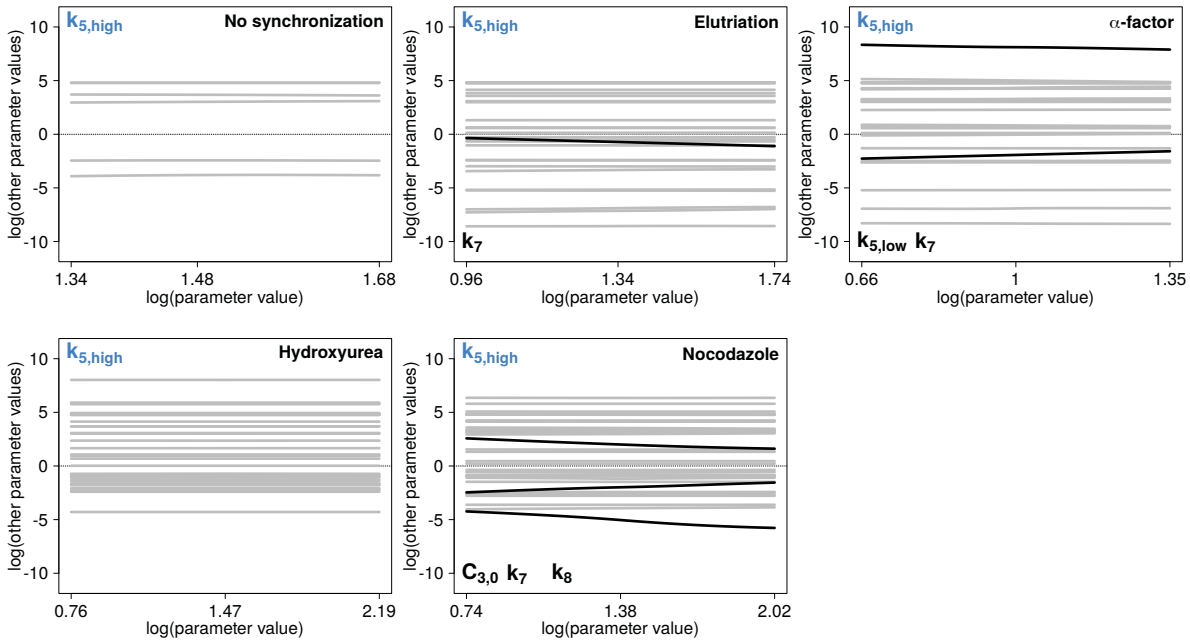

Figure 11.4: Parameter values along the profile likelihood of mRNA parameter $k_{5,high}$. In this figure, we show parameter values along the profile likelihood of the mRNA parameter $k_{5,high}$ which is related to mRNA species *CLN2* and is determined as being identifiable. We distinguish between parameter estimates of the mRNA optimization step calculated from smFISH data ("No synchronization") and parameter re-estimates of the protein optimization step calculated from Western blot data ("Elutriation", "α-factor", "Hydroxyurea" and "Nocodazole"). Parameter values are log-transformed. Gray and black lines indicate independent and dependent parameters, respectively. Dependent parameters are noted in the plot. A detailed plot description is given in Figure 6.2.

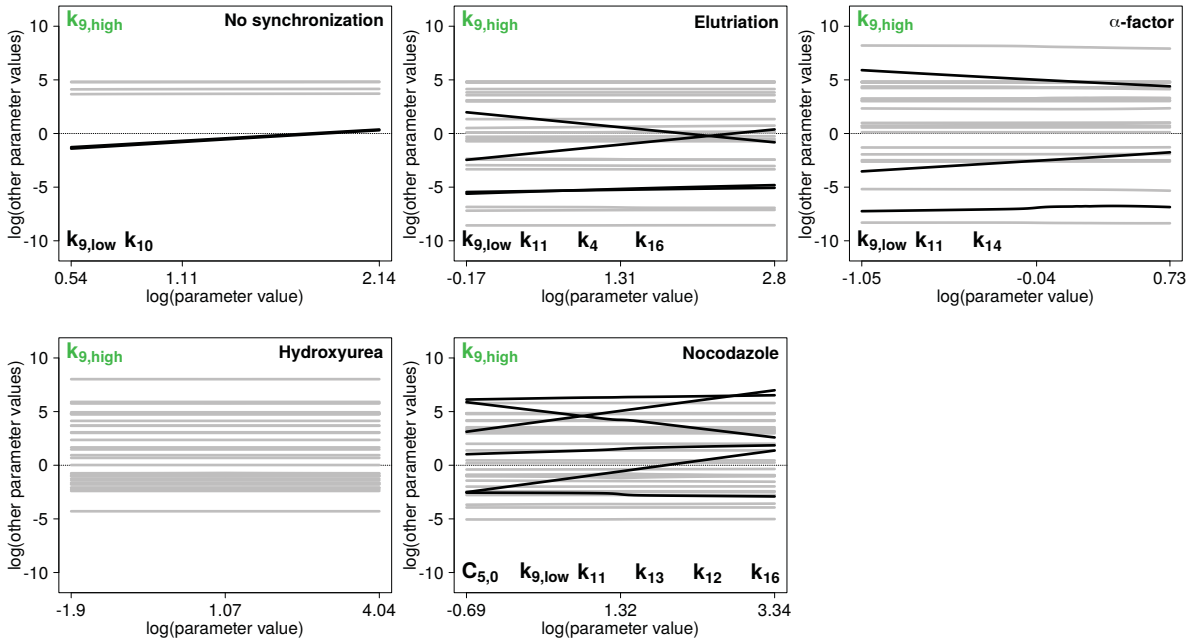

Figure 11.5: Parameter values along the profile likelihood of mRNA parameter $k_{9,high}$. In this figure, we show parameter values along the profile likelihood of the mRNA parameter $k_{9,high}$ which is related to mRNA species *CLB5* and is determined as being identifiable. We distinguish between parameter estimates of the mRNA optimization step calculated from smFISH data ("No synchronization") and parameter re-estimates of the protein optimization step calculated from Western blot data ("Elutriation", "α-factor", "Hydroxyurea" and "Nocodazole"). Parameter values are log-transformed. Gray and black lines indicate independent and dependent parameters, respectively. Dependent parameters are noted in the plot. A detailed plot description is given in Figure 6.2.

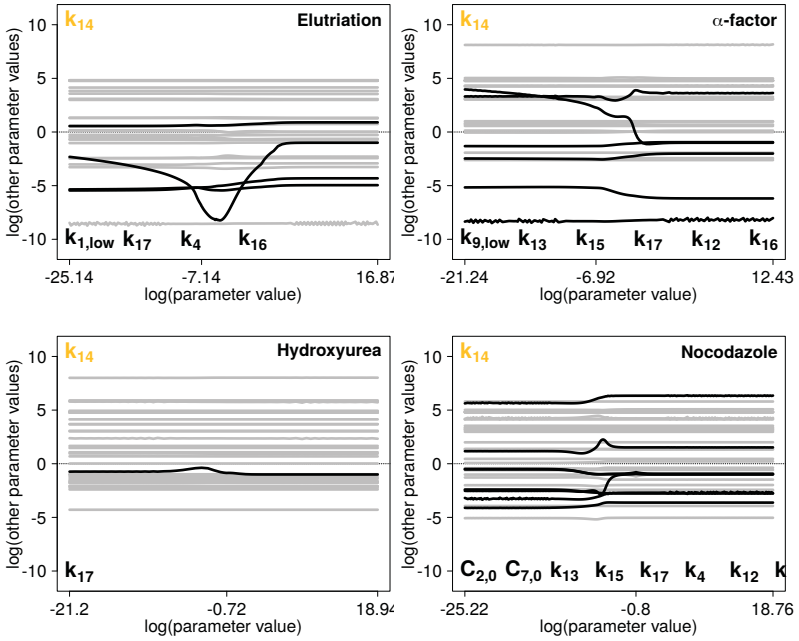

Figure 11.6: Parameter values along the profile likelihood of protein parameter k_{14}. In this figure, we show parameter values along the profile likelihood of the protein parameter k_{14} which is related to protein complex Sic1Clb5 and is determined as being structurally non-identifiable. We distinguish between parameter estimates calculated from Western blot data in the protein optimization step ("Elutriation", "α-factor", "Hydroxyurea" and "Nocodazole"). Parameter values are log-transformed. Gray and black lines indicate independent and dependent parameters, respectively. Dependent parameters are noted in the plot. A detailed plot description is given in Figure 6.2.

smaller molecule numbers in variable $C_9(t, \theta)$.

The practically non-identifiable parameter k_7 show variation in the same variables as parameter $k_{5,high}$. The difference is that variations are smaller and occur asymmetrically. Smaller values of parameter k_7 lead to larger molecule numbers and larger variations. It is not obvious that variations occur for variables $C_5(t, \theta)$ and $C_9(t, \theta)$ but we detected parameter dependencies for timing parameters of variable $C_3(t, \theta)$ to degradation rates (k_{10} and k_{16}) of variables $C_5(t, \theta)$ and $C_9(t, \theta)$.

Most variations occur for the structurally non-identifiable parameter k_{14}. Variables $C_3(t, \theta)$ and $C_4(t, \theta)$ are the only once not showing any variation. In contrast, trajectories of elutriated cells show these variations (see Figure E.24). Variations in variables $C_7(t, \theta)$, $C_8(t, \theta)$ and $C_9(t, \theta)$ indicate that we need measurements for protein complexes. Additionally, we can see how the variation in one variable transfers to other variables by the structure of the chemical reaction system, e.g. variation in variable $C_4(t, \theta)$ transfers to variable $C_3(t, \theta)$ and $C_7(t, \theta)$ to $C_8(t, \theta)$ (see Figures E.24 and E.25).

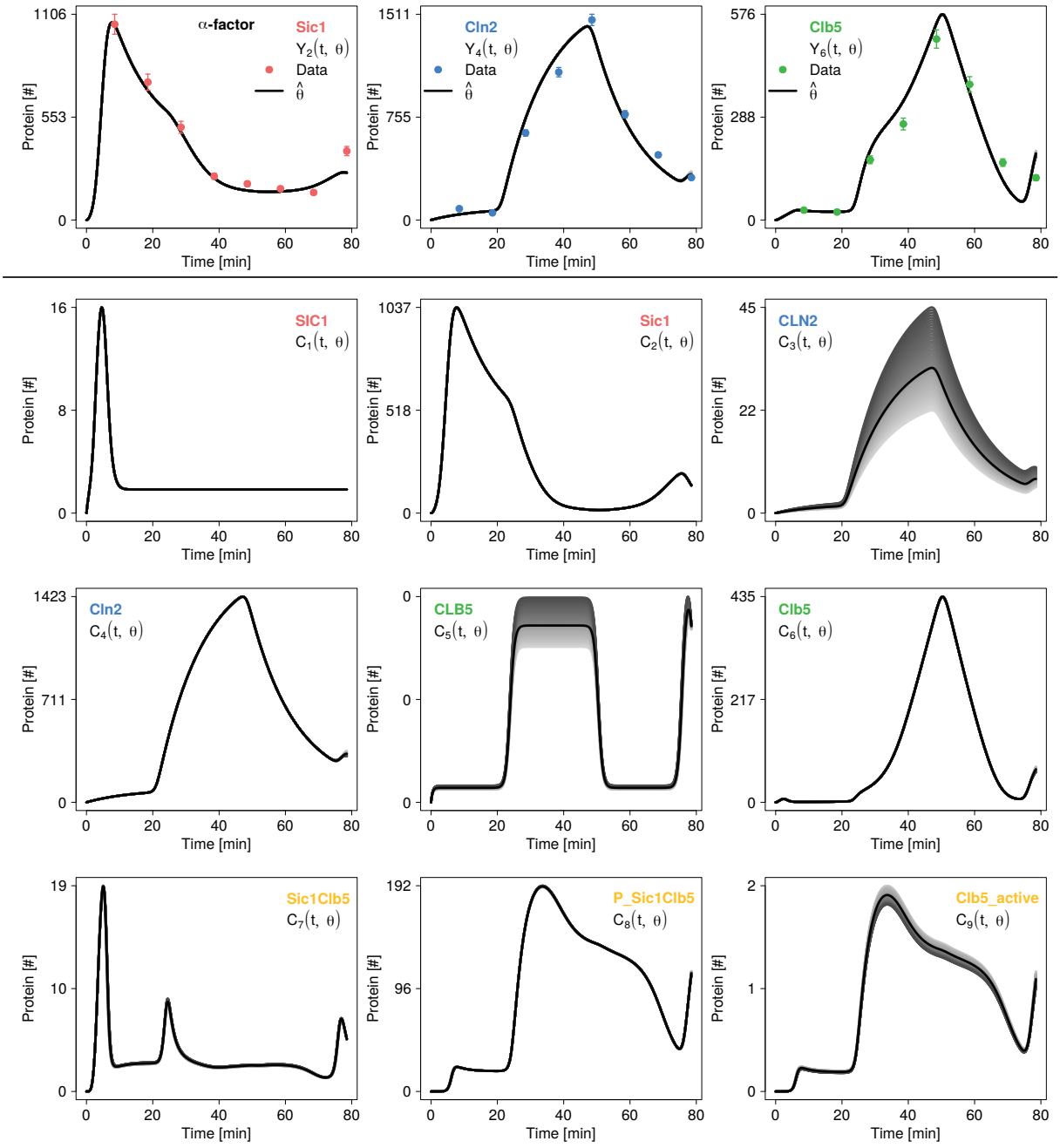

Figure 11.7: Trajectories along the profile likelihood of mRNA parameter $k_{5,high}$ re-estimated for α-factor synchronized cells in the protein optimization step. In this figure, we show trajectories of observables (above black horizontal line) and variables (below black horizontal line) along the profile likelihood of mRNA parameter $k_{5,high}$ re-estimated for α-factor synchronized cells in the protein optimization step. Observables are plotted together with the data (colored dots and error bars). Trajectories go from smallest (light gray) to largest (dark gray) parameter values. Trajectories for parameter optima are given as black lines. Red, blue and green species names refer to *SIC1*/Sic1, *CLN2*/Cln2 and *CLB5*/Clb5, respectively as in the data. Yellow represents a mixture of Sic1 and Clb5.

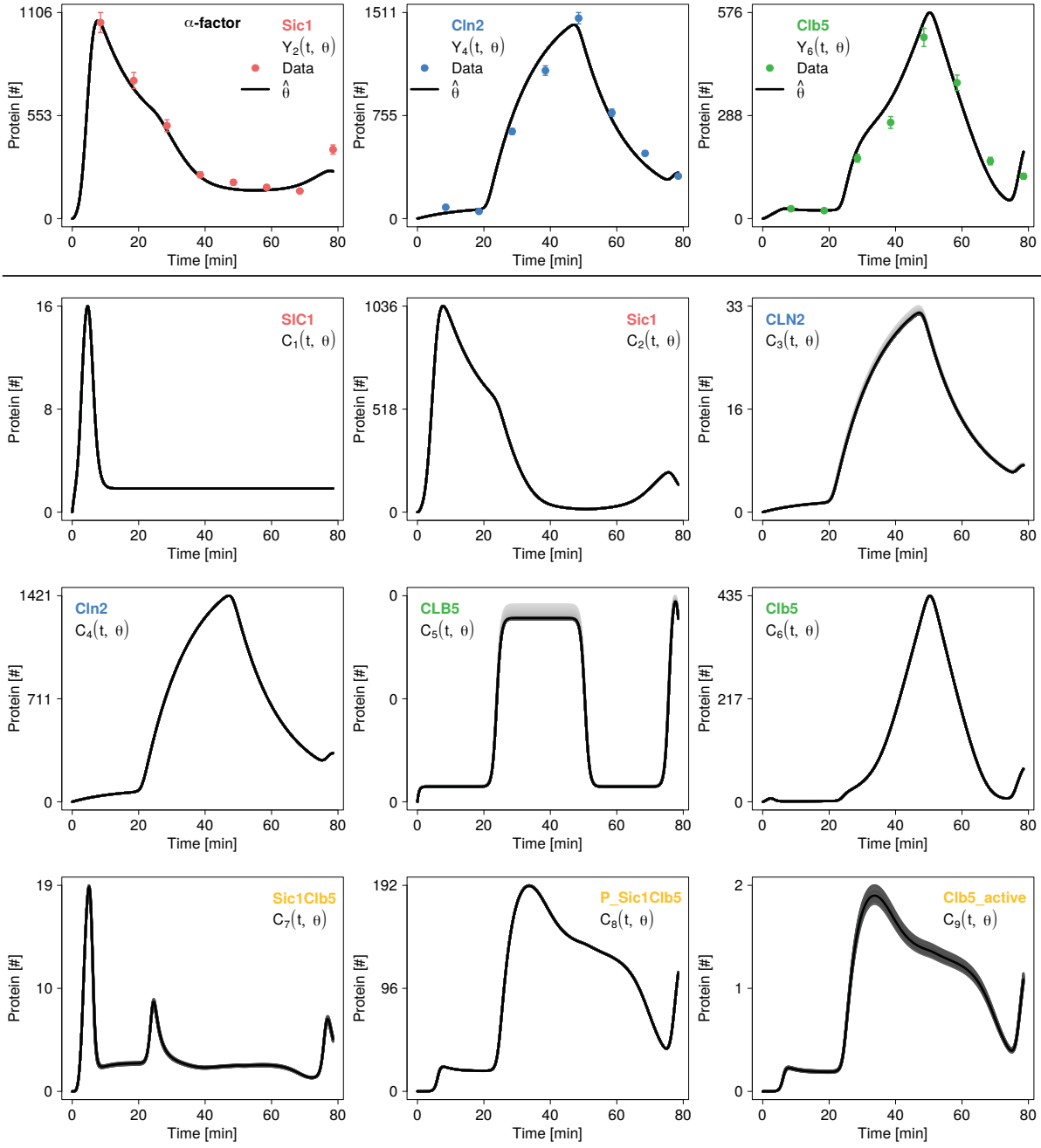

Figure 11.8: Trajectories along the profile likelihood of protein parameter k_7 estimated for α-factor synchronized cells in the protein optimization step. In this figure, we show trajectories of observables (above black horizontal line) and variables (below black horizontal line) along the profile likelihood of protein parameter k_7 estimated for α-factor synchronized cells in the protein optimization step. Observables are plotted together with the data (colored dots and error bars). Trajectories go from smallest (light gray) to largest (dark gray) parameter values. Trajectories for parameter optima are given as black lines. Red, blue and green species names refer to *SIC1*/Sic1, *CLN2*/Cln2 and *CLB5*/Clb5, respectively as in the data. Yellow represents a mixture of Sic1 and Clb5.

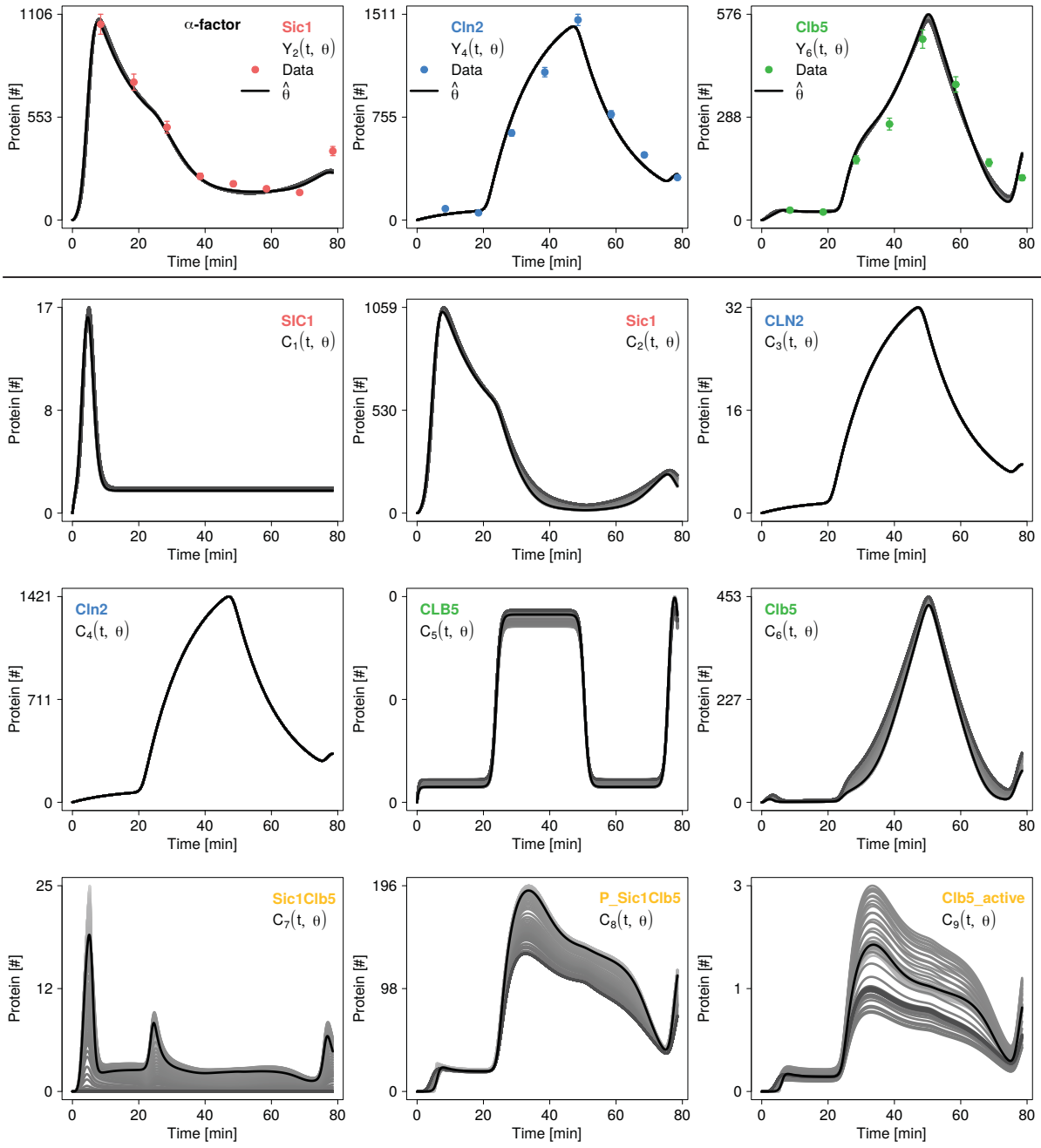

Figure 11.9: Trajectories along the profile likelihood of protein parameter k_{14} estimated for α-factor synchronized cells in the protein optimization step. In this figure, we show trajectories of observables (above black horizontal line) and variables (below black horizontal line) along the profile likelihood of protein parameter k_{14} estimated for α-factor synchronized cells in the protein optimization step. Observables are plotted together with the data (colored dots and error bars). Trajectories go from smallest (light gray) to largest (dark gray) parameter values. Trajectories for parameter optima are given as black lines. Red, blue and green species names refer to *SIC1*/Sic1, *CLN2*/Cln2 and *CLB5*/Clb5, respectively as in the data. Yellow represents a mixture of Sic1 and Clb5.

12. L_1 regularization

12.1 L_1 regularization improves identifiability of protein parameters

We apply L_1 regularization to protein parameters in the protein optimization step to figure out which protein parameters are not covered by Western blot data (see Section 7.2). In this L_1 regularization, we fixed mRNA parameters to re-estimated parameter values and the scaling factor to its estimate of the protein optimization step. Parameter fixation requires identifiable parameters. In Section 11.1, we have seen that this requirement is fulfilled for fixed parameters.

Identification of the optimal regularization strength $\hat{\eta}$ was successful (see Figure E.26 and Section 7.1 for calculation details). Synchronization by elutriation is the only synchronization method with fluctuations in the likelihood ratio $D_{protein}(\eta)$. We decided to set the optimal regularization strength to the value where $D_{protein}(\eta)$ finally exceed the 0.95 quantile of the $\chi^2(df)$ distribution. In parameter paths, we see that most non-selected parameters are going to zero quite fast (see Figure E.26).

Parameters k_{13}, k_{14} and k_{17} are removed (dark gray) in every synchronization method (see Table 12.1). These parameters were structurally non-identifiable for most synchronization methods (see Table 11.2) and refer to protein species Sic1Clb5 ($C_7(t, \theta)$) and P_Sic1Clb5 ($C_8(t, \theta)$). Compared to non-regularized parameter estimates, identifiability is essentially improved. Most selected parameters are identifiable (white) and structurally non-identifiabilities (orange) no longer exist (see Figures E.27, E.28, E.29 and E.30). Some practically non-identifiable parameters (light blue) become identifiable for smaller confidence levels.

Three parameters are not successfully selected (asterisk). Their profile likelihoods are compatible with zero (see Section 7.1 for calculation details). These parameters are candidates for supervised removal meaning that we can remove them as well. Parameter removement in the log-transformed parameter space is equivalent to parameter fixation to one in the non-transformed parameter space. Here, the total profile has only two contributions (data contribution and Gaussian prior). The mRNA prior is not needed if mRNA parameters are fixed and no longer re-estimated.

Reaction rates are uniquely set to zero. We re-introduced removed parameters one after the other and looked at the parameter values along the profile likelihood of the re-introduced parameter (see Section 7.1 for calculation details). The profile likelihood of the re-introduced parameter should be located around zero. If the line of another parameter horizontally crosses zero, this parameter can be removed instead of the actual removed parameter. Some initial values set to zero during L_1 regularization are exchangeable for synchronization by hydroxyurea and nocodazole (see Figures E.31 and E.32).

12.2 mRNA fold changes suggest synchronization specific smFISH measurements

We apply L_1 regularization in the protein optimization step to identify mRNA parameter differences between unsynchronized and synchronized cells (see Section 7.3). Therefore, we regularized mRNA fold changes $\tilde{r}$ which measure differences to mRNA parameters estimated from smFISH data in the mRNA optimization step. In this L_1 regularization, we fixed protein parameters to values received from applying L_1 regularization to protein parameters (see Section 12.1) to ensure parameter identifiability.

Identification of the optimal regularization strength $\hat{\eta}$ was successful but with more fluctuations in the likelihood ratio $D_{mRNA}(\eta)$ as seen for L_1 regularization of protein parameters (see Figure E.33 and Section 7.1 for calculation details). The likelihood ratio shows least fluctuations for Elutriation. We decided to set the optimal regularization strength to the value where

Par	Synchronization											
	Elutriation			α-factor			Hydroxyurea			Nocodazole		
	95%	90%	68%	95%	90%	68%	95%	90%	68%	95%	90%	68%
$C_{1,0}$												
$C_{2,0}$												
$C_{3,0}$												
$C_{4,0}$												
$C_{5,0}$												
$C_{6,0}$												
$C_{7,0}$												
$C_{8,0}$												
$C_{9,0}$												
k_3										*		
k_4												
k_7												
k_8				*			*					
k_{11}												
k_{12}												
k_{13}												
k_{14}												
k_{15}												
k_{16}												
k_{17}												

Removed | Not Estimated | Identifiable | Structural | Practical

Table 12.1: Protein parameter identifiabilities after L_1 regularization. This table shows identifiabilities for different confidence levels which are determined from total profiles of protein parameters estimated in the protein optimization step after applying L_1 regularization for elutriation, α-factor synchronization, hydroxyurea synchronization and nocodazole synchronization (see Figures E.27, E.28, E.29 and E.30). Structurally non-identifiable parameters are colored in orange and practical non-identifiable parameters are colored in light blue. Identifiable parameters are given in white. Initial values colored in light gray are not estimated. Cells colored in dark gray indicate parameters which gone to zero during L_1 regularization. Asterisks in the column of the 95% confidence region mark total profiles not successfully selected meaning that they are compatible with zero.

$D_{mRNA}(\eta)$ finally exceed the 0.95 quantile of the $\chi^2(df)$ distribution. We see in the parameter paths that fluctuations in the likelihood ratio does not change mRNA fold change selection (see Figure E.33.

The most mRNA fold changes are selected for α-factor synchronization and the least for Elutriation. Table 12.2 illustrates selected mRNA fold changes. mRNA fold changes larger than one indicate larger parameter values for synchronized compared to unsynchronized cells (orange). In contrast, mRNA fold changes smaller than one indicate smaller parameter values (light blue). Shared parameters among unsynchronized and synchronized cells are non-selected (white).

mRNA fold changes related to timing parameters are more frequently selected than reaction rates as seen in Section 10.2. Additionally, mRNA fold changes related to reaction rates are larger than those related to timing parameters and go well with larger confidence intervals (see Figure 10.2). Since we expected a relationship between Western blot data contributions to mRNA parameter (see Section 11.2) and determined mRNA fold changes, we counted the number of matches and performed a binomial test. Unfortunately, the number of matches is not significantly higher.

Hydroxyurea shows mRNA fold changes not successfully selected (asterisk) meaning that we could also set these mRNA fold changes to one (see Section 7.1 for calculation details). In this way, hydroxyurea would outperform elutriation. Estimated mRNA fold changes are identifiable (see Figures E.34, E.35, E.36 and E.37). Since we do not have prior knowledge about mRNA fold changes, total profiles have only contributions of the data and the Gaussian prior.

Selection of mRNA fold changes was not necessarily unique (see Section 7.1 for calculation details). In some cases, timing parameters can be exchanged by production rates (see Figures 12.1, E.38, E.39, E.40 and E.41). Thus, we cannot be sure that differences between unsynchronized and synchronized cells results from differences in cell cycle timing only. Measuring smFISH data for different synchronization methods would contribute to clarification by delivering information about positions of high transcription regions and mRNA production rates.

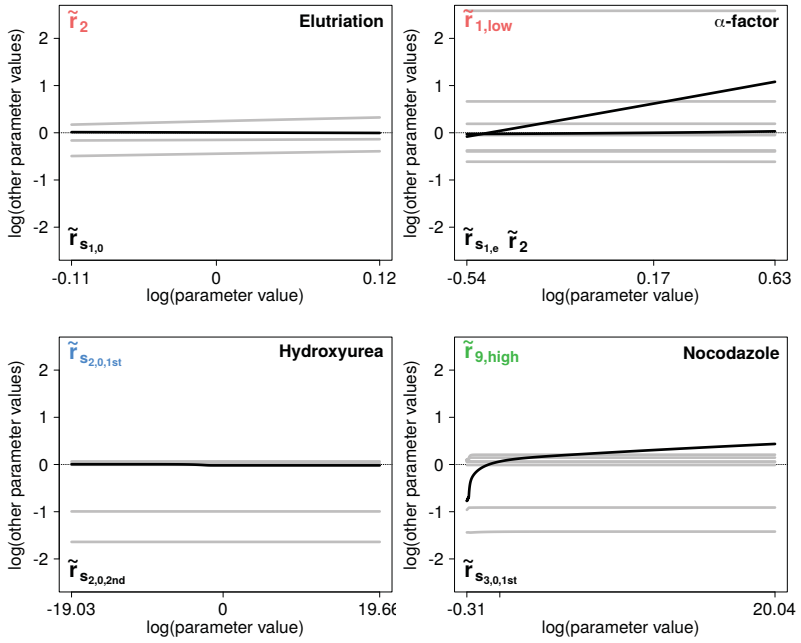

Figure 12.1: Parameter values along the profile likelihoods of L_1 removed mRNA fold changes. In this figure, we show parameter values along profile likelihoods of L_1 removed mRNA fold changes $\tilde{r}_1$, $\tilde{r}_{1,low}$, $\tilde{r}_{S_{2,0,1st}}$ and $\tilde{r}_{9,high}$ ("Elutriation", "α-factor", "Hydroxyurea" and "Nocodazole", respectively) which are subsequently re-introduced to test uniqueness of the solution. mRNA fold changes are log-transformed. Black lines indicate selected mRNA fold changes which are exchangeable with non-selected mRNA fold changes why non-selected mRNA fold changes are not uniquely determined to be zero. Exchangeable mRNA fold changes are noted in the plot. A detailed plot description is given in Figure 6.2.

Parameter	Synchronization			
	Elutriation	α-factor	Hydroxyurea	Nocodazole
$\tilde{r}_{S_{1,0}}$	1.01		1.00 *	1.07
$\tilde{r}_{S_{1,e}}$		0.99	1.00 *	1.06
$\tilde{r}_{S_{2,0,1st}}$				
$\tilde{r}_{S_{2,e,1st}}$		1.94		
$\tilde{r}_{S_{2,0,2nd}}$		0.95	0.98 *	0.99
$\tilde{r}_{S_{2,e,2nd}}$				
$\tilde{r}_{S_{3,0,1st}}$	0.86	0.69		0.73
$\tilde{r}_{S_{3,e,1st}}$		1.21		
$\tilde{r}_{S_{3,0,2nd}}$		0.96	1.02	1.16
$\tilde{r}_{S_{3,e,2nd}}$				1.22
$\tilde{r}_{1,low}$			1.07 *	
$\tilde{r}_{1,high}$				
$\tilde{r}_2$		1.57	0.19	0.24
$\tilde{r}_{5,low}$				
$\tilde{r}_{5,high}$		0.67		
$\tilde{r}_6$				
$\tilde{r}_{9,low}$	0.64	0.54		
$\tilde{r}_{9,high}$				
$\tilde{r}_{10}$	1.23	13.25	0.37	0.40

	$= \theta_\iota^{FISH}$		$> \theta_\iota^{FISH}$		$< \theta_\iota^{FISH}$

Table 12.2: mRNA fold changes. This table shows mRNA fold changes between mRNA parameters of unsynchronized and synchronized cells which are estimated by applying L_1 regularization. mRNA fold changes are rounded to the second decimal. mRNA parameters which are shared between unsynchronized and synchronized cells are colored in white. Parameter θ_ι^{FISH} is the corresponding mRNA parameter of unsynchronized cells and estimated in the mRNA optimization step. mRNA fold changes colored in orange show larger parameter values whereas mRNA fold changes colored in blue indicate smaller parameter values. Asterisks mark total profiles not successfully selected meaning that they are compatible with zero.

IV Discussion

13 Discussion **95**
13.1 Discussing method specific results
13.2 Regarding general interests of this study

14 Final statement **101**

13. Discussion

13.1 Discussing method specific results

In this study, we analyzed influences of different synchronization methods on the cell cycle and differences between unsynchronized and synchronized cells by using a stochastic modeling approach. We combined phase-resolved mRNA distributions of unsynchronized single cells and protein time courses of synchronized cell populations to estimate model parameters and to predict synchronization specific mRNA dynamics. Parameter estimation is based on methods developed for deterministic model systems and performed in a 2-step-optimization in which we differentiate between mRNA and protein level.

Experimental data

In the mRNA optimization step, we assume Poisson distributed mRNA distributions to use the analytical solution of the CME which is a Poisson distribution parameterized by the respective RRE of the birth-death process for a Poisson initial distribution. We showed that assuming Poisson distributed mRNA distributions is a good approximation to estimate mRNA parameters even if only a few mRNA distributions are significantly Poisson distributed. Poisson distributions have a limited capacity to cover the distributional spread of mRNA distributions. The negative binomial distribution is an example of a discrete probability density function in which mean and variance are not equal. If we use those probability density functions, we can probably better describe mRNA distributions but the relationship between the analytical solution of the CME and the RRE is no longer existing. However, this relationship is needed to simulate phase-resolved mRNA distributions used in parameter estimation.

In [118], authors suggest that mRNA distributions over the whole cell cycle are better described by a two component compared to a single component Poisson distribution. In this case, ODE constrained mixture models can be applied to simulate overall distributions based on two subpopulations with different means [119, 120]. We showed that mRNA distributions over the whole cell cycle follow neither a single component nor a two component Poisson distribution.

Gene expression in yeast is dominated by extrinsic noise [45]. In the given smFISH data, extrinsic noise cannot be analyzed because individual mRNA species are measured per cell. Recently, a new method for multi-gene detection on the single cell level was presented in [121] and already established in our lab. Now, we can count up to three mRNA species per cell and analyze correlations between them.

Data pre-processing of Western blot data requires repetitive measurements. We measured different numbers of technical replicates per synchronization method and pre-processed them by using the R package **blotIt2**. This tool is easy to use and leads to successful parameter estimation in the protein optimization step. Unfortunately, its documentation is more or less missing. Compared to error estimation as part of the optimization problem, data pre-processing results in larger variances of parameter estimates [77]. Technical replicates do not represent biological variability. Biological variability reflects inter-individual and inter-experimental variability and leads to more general results [76]. People use time warping algorithms to map protein time courses of different measurements to increase the number of replicates [122, 123]. If we use these algorithms for our synchronization specific protein time courses, biological variability can be represented but we can no longer analyze synchronization effects on the cell cyle.

Biological data show multiplicative and log-normally distributed noise [77, 79, 80]. The application of additive normal noise models requires a log-transformation of the data. We used non-transformed data in the protein optimization step for two reasons. First, log-transformed and non-transformed pre-processed Western blot data were similar. Second, log-transformation was not possible for a few data point due to negative band intensities.

Mathematical model

Simulated mRNA distributions are in good agreement with measured mRNA distributions. However, simulated mRNA distributions have certain characteristics which make a comparison with measured mRNA distributions difficult. Dependent on the position of the high transcription region in relation to cell cycle phases, simulated mRNA distributions are far from being Poisson distributed. In cell cycle phases where mRNA numbers are in- or decreasing, mRNA distributions are skew. In contrast, mRNA distributions show two accumulation points in cell cycle phases where transitions between low and high transcription regions take place. Further, simulated mRNA distributions show an enhanced occurrence of zero mRNA molecules for larger mean values. Measured mRNA distributions are neither skew nor do they show two accumulation points or a strong occurrence of zero mRNA molecules for large mean values. At this point, it becomes clear that (i) a statistical test will always lead to significant differences between measured and simulated mRNA distributions and (ii) predicted synchronization specific mRNA dynamics have to be validated by measuring smFISH data in synchronized cells.

In addition, we used equidistantly scaled cell cycle phase lengths determined for unsynchronized cells to simulate phase-resolved mRNA distributions of different synchronizations but we do not know how different these cell cycle phases are in synchronized cell populations. Furthermore, transitions between low and high transcription regions occurs at the same time for every simulated cell and, therefore, do not allow for transition variability. In [124], authors show that the G1 to S phase transition is dependent on cell size and occur later in daughter cells compared to mother cells. We do not differentiate between mother and daughter cells or take cell size into account but fixed transition times are still unreliable.

Simulated protein time courses are in good agreement with measured time courses. In some time windows, additional data points are required for clarification, e.g. Clb5 in nocodazole synchronized cells. Additionally, mismatches between measured and simulated time courses can be a result of an oversimplified RRE. We included elementary reactions in the stochastic model formulation wherefore the respective RRE includes mass action kinetics only. However, most published deterministic models of the cell cycle include more complex kinetics [12, 125, 126, 127, 128].

Regarding mass action kinetics only, the number of estimable parameters is as small as possible and, therefore, we reduce the number of non-identifiable parameters. Moreover, mass action kinetics in the deterministic model formulation have a direct link to the stochastic model formulation. This is why we can estimate deterministic rate constants to determine stochastic rate constants as also done in [5]. In complex cell cycle models, parameters are adjusted to experimental data and not estimated or the majority of parameters have to be non-identifiable. Unfortunately, authors often do not provide necessary informations about their model parameters.

Parameter estimation

In [77], authors recommend latin hypercube sampling in combination with a trust region optimization to estimate parameters of ODE systems from time course data. In the mRNA optimization step, we got a better performance for the combination of latin hypercube sampling and a line search optimization. The likelihood waterfall plots of the trust region optimization constantly increased and did not show any clear optima. This is probably because we used a different likelihood function.

Further, authors recommend to solve sensitivity equations to determine model sensitivities needed to calculate the gradient of the objective function to prevent numerical errors. We got better results by using the finite difference approximation instead. Note, we chose a step control where numerical errors are still small. Numerical errors can also be avoided by using complex-step derivatives instead of finite differences [63, 129, 130]. Unfortunately, calculation times grew to infinity and we did not use them anymore. For large model systems, adjoint sensitivities are utilized as well [97].

The performance of the protein optimization step is not satisfactory at all. The global optimum is not well represented in every synchronization method. There are generally two possible explainations. First, a multi-start local optimization is not the preferable stochastic

optimization method. In [60], authors summarize a number of stochastic optimization methods we can try. Second, the whole optimization setup is insufficient. In this case, it is more difficult to improve optimization because there are many ways to intervene in the overall system. A possibility is that the parameter space was not well represented by the sampling method (Gaussian sampling) or that we performed too little optimization runs. In the same way, the observation function can be wrong [58, 104].

We implemented induction of low and high transcription regions by time dependent mRNA production rates. In a different ansatz, we can implement signals by trigger functions [131] and check how optimization change if timing is not directly coupled to mRNA production rates. A further change can be to use non-normalized Western blot data and use informations given by the PaxDb database as prior knowledge for initial values. We normalized mean values of protein time courses to absolut protein numbers per cell. In our opinion, the used normalization is most suited for time courses with different maximal protein levels. Normalization enables to combine relative protein numbers with absolut mRNA numbers.

A big issue is that we want to figure out differences between synchronization methods but we do not have a statistical test to analyze differences between parameter estimates. We calculated confidence intervals from profile likelihoods but we cannot calculate measures needed for statistical testing, e.g. variances used in an ANOVA. We have a single parameter estimate but no underlying parameter distribution. There are some ideas how to test significant differences if confidence intervals are available only [117, 132, 133]. However, looking for non-overlapping confidence intervals is the only convincing method to judge significant differences.

Identifiability analysis

The chemical reaction system treated in this study is taken from [50] and adjusted to our problem statement. In the original system, authors fixed some parameter values according to literature values and estimated others. As in our study, authors could not present a fully identifiable mathematical model and, therefore, model predictions are questionable. This is why we apllied L_1 regularization for model reduction.

We showed that non-availability of experimental data for protein complexes is the main problem for parameter estimation. New measurements would change the observation function and prevent overlaps between observables. As long as we cannot remove these overlaps, the system will never be fully identifiable and predictable. At present, protein complexes cannot be measured. They are extremely short-lived and low abundant. There are no antibodies to apply methods as co-immunoprecipitation presented in [78].

In the protein optimization step we introduced prior knowledge for mRNA parameters. These mRNA priors contribute to the total profile of each estimated parameter and, therefore, influences identifiability of each parameter as well. mRNA priors are reasonable for mRNA parameters and questionable for protein parameters. Therefore, we analyzed the mRNA prior contribution to the total profile of protein parameters. Another possibility is to include Western blot data and Gaussian priors in the identifiability analysis of protein parameters only. We decided to include mRNA priors because they were essential for a successful fitting of the full model to Western blot data. Profile likelihoods of the reduced model have no mRNA prior contributions.

We determined mRNA prior probability density functions by using parameter estimates and confidence intervals determined in the mRNA optimization step. Thus, mRNA parameters can be slightly adjusted in the protein optimization step. Different mRNA priors are conceivable. A strategy for systematical testing of different mRNA prios is to use parameter estimates of the mRNA optimization step only, set variances to a common value and gradually increase this value.

We performed identifiability analysis after parameter estimation. A more efficient strategy is to perform an additional identifiability analysis before estimating model parameter to fix structurally non-identifiable parameters a priori [112]. Using symmetry detection which is implemented in the R package **dMod** [63], we found that for given observables of the protein optimization step three symmetry groups remain. If we fix respective parameter values (protein production rates) in the full model, identifiability can be further improved.

Interpreting parameter dependencies is difficult because detected dependencies are not equiv-

alent to what is presented in the literature. In [110], authors found parameter dependencies between structurally non-identifiable parameters only. We found parameter dependencies between any combination of identifiable, practically non-identifiable and structurally non-identifiable parameters. Parameter dependencies shown in [110] are unique. Lines between independent parameters are straight and between dependent parameters in- or decreasing. In our plots, lines are not always unique. For this reason, we decided for the selection criterion that variances have to be smaller than 0.01 for independent parameters.

In the same publication, authors analyzed variations in trajectories of observables and variables. They stated that variations in observables arise along profile likelihoods of practically non-identifiable parameters only and variations in variables along practically and structurally non-identifiable parameters. We found variations in observables along profile likelihoods of structurally non-identifiable parameters and variations in variables along identifiable parameters as well. In contrast to the author's parameter estimation, we used mRNA priors in the protein optimization step which are probably the main source for differences in parameter dependencies and variations in trajectories. Further reasons are convergence problems as well as asymmetric and broad profile likelihoods.

L_1 regularization

We applied L_1 regularization to (i) find shared mRNA parameters between unsynchronized and synchronized cells and to (ii) remove non-identifiabilities due to model reduction. The first idea is taken from [113] and delivered similar results as analyzing Western blot data contributions to total profiles of re-estimated mRNA parameters. Not all mRNA fold changes were uniquely selected. Some timing parameters are exchangeable with mRNA production rates. smFISH data for different synchronization methods are needed for further analysis. We cannot determine shared protein parameter values due to the lack of protein measurements in unsynchronized cells. A possibility is to look for shared mRNA and protein parameters between different synchronizations only.

Regarding model reduction, we regularized protein parameters only. L_1 regularization can also be applied to the complete parameter set. In our view, it is not useful to regularize mRNA parameters on the basis of Western blot data. We showed that identifiability of protein parameters is essentially improved in the reduced model. Other methods can be used for model reduction. In [116], authors present a flow-chart which represents a strategy to reduce model complexity by systematically analyzing profile likelihoods and, therefore, avoiding the need of new measurements. Structurally non-identifiable parameters are removed by fixation and practically non-identifiable parameters by reaction removal, algebraic substitutions or context-specific reductions.

Model reduction improve identifiability and prevent new experimental data. Nevertheless, new experimental data are mandatory regarding a specific biological question. Our mathematical model was established to analyze the role of protein complexes in the G1 to S phase transition. As seen in this study, new experimental data are required to perform this analysis. There are techniques to efficiently plan new experiments. A possibility is to perform Monte Carlo simulations to test the benefit of new experimental data [58, 76].

In both L_1 regularizations we tested a range of different regularization strengths. We used a sampling rate of approximately 0.03. This number was a trade-off between the overall calculation time and the smallest possible distance to adjacent values. A smaller sampling rate is always desirable. However, exorbitantly small sampling rates let calculation times explode without considerably influencing the optimal regularization strength.

13.2 Regarding general interests of this study

Biological questions

In this study, we showed that synchronization influences cell cycle behavior. Our analysis suggests that cell cycle timing is mainly responsible for synchronization specific mRNA dynamics. Nevertheless, cell cycle timing and gene transcription differ between different synchronization

methods and between unsynchronized and synchronized cells. Cell cycle timing is mostly affected in α-factor synchronized cells. The first high transcription regions of *CLN2* and *CLB5* span approximately two cell cycle phases. In contrast, gene transcription is mostly affected in hydroxyurea synchronized cells. The number of mRNA molecules is about twice as large as in unsynchronized cells, especially for *SIC1*, whereas high transcription regions are almost unchanged. Nocodazole synchronized cells show changes in high transcription regions and gene transcription. In total, elutriated cells are closest to unsynchronized cells. High transcription regions are slightly prolonged compared to unsynchronized cells and gene transcription is similar.

Since we expect most synchronization effects in the first cell cycle, subsequent cell cycle passages can be used to analyze "normal" cell cycle behavior [28]. However, it is not sure that cellular processes are completely recovered from the synchronization procedure and how far desynchronization is proceeded. In our measurements, protein oscillations persist for three cell cycle passages for nocodazole only. Elutriated cells already desynchronize in the first cell cycle and do not really show any protein oscillation. We suggest to model desynchronization of different synchronization methods which can be used for "trend elimination" as part of data pre-processing. Otherwise, the development of new or advanced synchronization methods with less intervention in cell cycle behaviour are required. A synchronization method comparable to centrifugal elutriation with less intervention is presented in [36].

Mathematical modeling tasks

In this study, we parameterized a stochastic model by using the respective deterministic model representation and a maximum likelihood approach. Therefore, we combined qualitatively different data types in a 2-step-optimization. In the mRNA optimization step, we estimated mRNA parameters from smFISH data which are phase-resolved mRNA distributions. In the protein optimization step, we re-estimated mRNA parameters and estimated protein parameters from Western blot data which are protein time courses. Both optimization steps are linked by (i) mRNA prior probability density functions calculated from mRNA parameter estimates of the mRNA optimization step and then used in the protein optimization step and (ii) a scaling factor introduced in the protein optimization step to transcription start and end times to overcome differences in cell division times.

The definition of a combined likelihood function for mRNA (Poisson error model) and protein level (Gaussian error model) is possible but also more challenging. First, an appropriate weighting for individual likelihood functions has to be found. Second, a common optimization algorithm has to be used. So far, we used a line search algorithm in the mRNA optimization step and a trust region algorithm in the protein optimization step.

Vision

In this study, we showed two important things. First, it is possible to use available data to make predictions based on mathematical modeling and to better understand biological processes. Second, it is not possible to avoid new experimental data if information is missing in available data and if predictions are not yet validated. Model reduction is a suited methodology to deal with missing information. As an example, we do not have information about protein complexes which is why we can fix related parameters to improve the overall parameter identifiability. In contrast, there is no possibility to prevent model validation, otherwise predictions are not reliable.

14. Final statement

In this study, we presented a systematic analysis of synchronization effects on the yeast cell cycle. We showed how different synchronization methods influences the cell cycle behavior and which differences exist between unsynchronized and synchronized cells. Therefore, we brought together mRNA measurements of unsynchronized single cells and protein measurements of synchronized cell populations.

We used mathematical modeling to predict mRNA dynamics of synchronized cells based on their protein dynamics. For this purpose, we combined qualitatively different data types, distribution and time series data, to parameterize a stochastic model system by using optimization methods developed for deterministic model systems.

Our analysis give rise to a number of ambiguities but still form the basis for further analysis. There is a strong evidence that it is not possible to derive general cell cycle behavior from synchronized cell populations. Methods presented in this study provide a fundament for analyzing synchronization effects also in other eukaryotic cells.

V

Appendix

A Symbol directory 105

B Abbreviations 111

C Definitions 113
C.1 Definition of time points t_{ln}
C.2 Definition of the number of cells $\mathcal{Z}_{kl}$
C.3 Definition of the number of mRNA molecules A_{ka}

D Calculations 115
D.1 Calculation of pre-processed data by using blotIt2
D.2 Calculation of the time evolution of the expected state
D.3 Calculation of the time until next reaction for time dependent stochastic rates
D.4 Calculation of the stationary distribution of a birth-death process
D.5 Calculation of the analytical solution of the CME for a birth-death process with a Poisson initial distribution
D.6 Calculation of the gradient of the objective function in the mRNA optimization step
D.7 Calculation of the gradient of the objective function in the protein optimization step
D.8 Calculation of time dependent mRNA production rates by using Fermi-Dirac distributions
D.9 Calculation of p-values in a χ^2 goodness of fit test
D.10 Latin hypercube sampling versus Gaussian sampling

E Figures 121

F Tables 155

A. Symbol directory

Symbol	Description
t	System time
τ	Time until next reaction
$\tilde{\tau}$	Leap time
dt, $d\tau$	Infinitesimal time increments
Δt	Finite time step
T_{cycle}	Cell division time
t_{ln} with $n = 1, ..., nT$	Time points in a cell cycle phase
t_n with $n = 1, ..., nT$	Time points in the whole cell cycle
h	Step control in finite difference approximation
$p_{MB}(\boldsymbol{v})$	Maxwell-Boltzmann PDF
$\boldsymbol{v} = (v_x, v_y, v_z)$	Velocity vector
k_B	Boltzmann constant
T	Absolute temperature
m	Molecular mass
n_A	Avogadro's constant
$V = vol \cdot dm^3$	Systems volume
Ω	$n_A \cdot vol$
N	Number of species
M	Number of reactions
$nPar$	Number of (estimated) parameters
$nPar_0$	Number of zero valued parameters associated with L_1 regularization
nP	Number of cell cycle phases
$nRep$	Number of replicates
nT	Number of time points
nR	Total number of measured mRNA molecules
nB	Number of bins
$nObs$	Number of observables
n_t	Number of time points within a time interval
N_{LHS}	Number of samples associated with latin hypercube sampling
S_i with $i = 1, ..., N$	Species
$X_i(t)$ with $i = 1, ..., N$	Number of molecules
$\boldsymbol{X}(t) = (X_1(t), ..., X_N(t))$	Discrete system state vector
$\langle \boldsymbol{X}(t) \rangle$	Expected state of $\boldsymbol{X}(t)$
$\langle \boldsymbol{X}^\star \rangle$	Expected state of $\boldsymbol{X}(t)$ in steady state
$\mathcal{R}$	Number of realizations
$X_{ir}^{SSA}(t)$ with $r = 1, ..., \mathcal{R}$	Numerical realization of $X_i(t)$
$\overline{X_i^{SSA}(t)}$	Average over $\mathcal{R}$ realizations of $X_{ir}^{SSA}(t)$
$\boldsymbol{X}^{SSA}(t)$	Numerical realization of $\boldsymbol{X}(t)$
$\overline{\boldsymbol{X}^{SSA}(t)}$	Average over $\mathcal{R}$ realizations of $\boldsymbol{X}^{SSA}(t)$
$\boldsymbol{x} = (x_1, ..., x_N)$	Acceptable values of $\boldsymbol{X}(t)$
$\boldsymbol{Z}(t) = (Z_1(t), ..., Z_N(t))$	Discrete intermediate system state vector

$\boldsymbol{z} = (z_1, ..., z_N)$	Acceptable values of $\boldsymbol{Z}(t)$
$C_i(t)$	Species concentration
$\boldsymbol{C}(t) = (C_1(t), ..., C_N(t))$	Continuous system state vector
$\boldsymbol{C}(t, \theta) = (C_1(t, \theta), ..., C_N(t, \theta))$	Continuous system state vector regarding an explicit dependency on θ
$\boldsymbol{C}_0 = (C_{1,0}, ..., C_{N,0})$	Initial system state of $\boldsymbol{C}(t, \theta)$
$\boldsymbol{f}(\boldsymbol{C}(t, \theta), \theta))$	Function describing the temporal evolution of $\boldsymbol{C}(t, \theta)$
$\boldsymbol{g}(\boldsymbol{C}(t, \theta), \theta))$	Observation function
$Y_k(t, \theta)$	Observable
$\boldsymbol{Y}(t, \theta) = (Y_1(t, \theta), ..., Y_{nObs}(t, \theta))$	Observable vector
y_k	Experimental data
$\boldsymbol{y} = (y_1, ..., y_{nObs})$	List of experimental data
R_j with $j = 1, ..., M$	Reaction
$\boldsymbol{\nu}_j = (\nu_{1j}, ..., \nu_{Nj})$	State-change vector
$\boldsymbol{\nu}$	Stoichiometric matrix
ϵ	$N \times N$ identity matrix
$s_1(t)$, $s_2(t)$, $s_3(t)$	Transcription signals
$t_{s_{1,0}}$, $t_{s_{2,0,1st}}$, $t_{s_{2,0,2nd}}$, $t_{s_{3,0,1st}}$, $t_{s_{3,0},2nd}$	Transcription start times
$t_{s_{1,0}}$, $t_{s_{2,e,1st}}$, $t_{s_{2,e,2nd}}$, $t_{s_{3,e,1st}}$, $t_{s_{3,e},2nd}$	Transcription end times
c_j	Stochastic rate constant
$c_{j,high}$, $c_{j,low}$	Stochastic rate constants for low and high transcription regions
$c_j(t)$	Time dependent stochastic rate
k_j	Deterministic rate constant
$k_{j,high}$, $k_{j,low}$	Deterministic rate constants for low and high transcription regions
$k_j(t)$	Time dependent deterministic rate
κ	Constant in the deviation of the stationary distribution of a birth-death process
$y_{k,(rep}(t_n)$	Measured value of observable y_k in replicate rep at time t_n
$f_{k,rep}(\theta^{pre}, t_n)$	Response of the biological system at time t_n
$\tilde{y}_k(t_n)$	Time course parameter
s_{rep}	Scaling parameter
$\theta^{pre} = (\tilde{y}_k(t_{n=1}), ..., \tilde{y}_k(t_{n=nT}),$ $s_{rep=1}, ..., s_{rep=nRep})$	Pre-processing parameter vector
$\theta = (k_1, ..., k_M, C_{1,0}, ..., C_{N,0})$	Parameter estimation vector
θ_ι	Specific parameter in θ
$\boldsymbol{e}_\iota$	ι-th identity vector associated with θ_ι
$\hat{\theta}$	Estimate of θ
θ^r	mRNA parameters
θ^p	Protein parameters
θ^k	Parameters of observable Y_k
$\hat{\theta}^k$	Estimate of θ^k
$\theta^{FISH} = (\theta^{k=1}, \theta^{k=3}, \theta^{k=5})$	mRNA parameters estimated from smFISH data
$\hat{\theta}^{PL}$	Parameter estimate of $PL(\boldsymbol{y}\|\theta_\iota)$
$\hat{\theta}^{bias}$	Biased estimate of θ
θ^η	η specific parameters
$\hat{\theta}^\eta$	Estimate of θ^η
$\theta^{p,\eta}$	η specific protein parameters

$\hat{\theta}^{p,\eta}$	Estimate of $\theta^{p,\eta}$
θ_ι^*, θ_ι^{r*}, θ^*	Known/true parameter values
$\tilde{r}$	mRNA fold changes
$\tilde{r}^\eta$	η specific mRNA fold changes
$\hat{\tilde{r}}^\eta$	Estimate of $\tilde{r}^\eta$
$\|\theta\|_\rho$	ρ-norm of vector θ
$a_j(\boldsymbol{X}(t))$	Propensity function
$\boldsymbol{a}(\boldsymbol{X}(t))$	Propensity vector
$a_j(\boldsymbol{X}(t,\theta),\theta)$	Propensity function regarding an explicit dependency on θ
$\boldsymbol{a}(\boldsymbol{X}(t,\theta),\theta)$	Propensity vector with an explicit dependency on θ
$h_j(\boldsymbol{X}(t))$	Combinatorial function
$P(\boldsymbol{x},t\|\boldsymbol{x}_0,t_0)$	Conditional PDF associated with the CME
$P^\star(\boldsymbol{x},t)$	Stationary distribution of the CME
$P(\boldsymbol{x}-\boldsymbol{\nu}_j,t\|\boldsymbol{x}_0,t_0)$	Conditional PDF associated with the CME
$P(\boldsymbol{x},t+dt;\boldsymbol{z}_j,t;\boldsymbol{x}_0,t_0)$	Joint PDF associated with the deviation of the CKE
$P(\boldsymbol{x}_0,t_0)$	PDF associated with the deviation of the CKE
$P(\boldsymbol{z}_j,t\|\boldsymbol{x}_0,t_0)$	Conditional PDF associated with the deviation of the CKE
$P(\boldsymbol{x},t+dt\|\boldsymbol{z}_j,t;\boldsymbol{x}_0,t_0)$	Conditional PDF associated with the deviation of the CKE
$P(\boldsymbol{x},t+dt\|\boldsymbol{z}_j,t)$	Markov process assumption for $P(\boldsymbol{x},t+dt\|\boldsymbol{z}_j,t;\boldsymbol{x}_0,t_0)$
$P(\boldsymbol{x},t+dt\|\boldsymbol{x}_0,t_0)$	Conditional PDF associated with the CKE
$P(\boldsymbol{x},t+dt\|\boldsymbol{x},t)$	Conditional PDF associated with the CKE
$P(\boldsymbol{x},t+dt\|\boldsymbol{x}-\boldsymbol{\nu}_j,t)$	Conditional PDF associated with the CKE
$p(\tau;j\|\boldsymbol{x},t)$	Reaction probability density function associated with the SSA
$P_{01}(\tau\|\boldsymbol{x},t)$, $P_{02}(j\|\boldsymbol{x},t)$	Conditional PDFs associated $p(\tau;j\|\boldsymbol{x},t)$ for τ and j being independent
$P_{01}(\tau+d\tau\|\boldsymbol{x},t)$	Conditional PDF associated with the deviation of $P_{01}(\tau\|\boldsymbol{x},t)$
$P_1(\tau\|\boldsymbol{x},t)$, $P_2(j\|\boldsymbol{x},t;\tau)$	Conditional PDFs associated $p(\tau;j\|\boldsymbol{x},t)$ for τ and j being dependent
$f_{FD}(t,\vartheta,\beta)$	Fermi-Dirac distribution
$f_{1-FD}(t,\vartheta,\beta)$	Mirrored Fermi-Dirac distribution
β, ϑ	Parameters associated with f_{FD}
$f_U(\xi)$	Uniform distribution (PDF)
$F_U(\xi)$	CDF of $f_U(\xi)$
$f_{Exp}(\tau)$	Exponential distribution (PDF)
$F_{Exp}(\tau)$	CDF of $f_{Exp}(\tau)$
$f_{point}(j')$	Point probability of the next reaction index in SSA (discrete PDF)
$F_{point}(j)$	CDF of $f_{point}(j')$
$\chi^2(df)$	χ^2 distribution (PDF)
$F_{\chi^2}(\chi_t^2)$	CDF of $\chi^2(df)$
χ_t^2	Test statistic of the χ^2 goodness of fit test
χ_c^2	Critical value of the χ^2 goodness of fit test
$\chi^2_{df,(1-\alpha)}$	$(1-\alpha)$ quantile of $\chi^2(df)$ distribution
$\mathcal{P}_j(\lambda)$, $\mathcal{P}(x,\lambda)$, $\mathcal{P}(\lambda)$	Poisson distribution (discrete PDF)
$\mathcal{P}_{TP}(\lambda)$	Time point Poisson distribution (discrete PDF)
$\mathcal{P}_{TI}(\lambda)$	Time interval Poisson distribution (discrete PDF)
$\boldsymbol{\mathcal{P}}$	Matrix of Poisson distributions associated with $L_{mRNA}(y_k\|\theta)$ and $\ell_{mRNA}(y_k\|\theta)$

$\mathcal{P}_{an}$	Elements of $\boldsymbol{\mathcal{P}}$
$\mathcal{N}_j(\mu, \sigma^2)$	Normal distribution (PDF)
$\mathcal{N}_j(0, 1)$	Standardized normal distribution (PDF)
$W_j(t)$	Brownian motion
$p_{Gauss}(\theta)$	Gaussian prior distribution (PDF)
$p_{mRNA}(\theta^r)$	mRNA prior distribution (PDF)
$\phi(y_k\|\theta)$	Conditional PDF associated with $L(y_k\|\theta)$
$\boldsymbol{\varepsilon}$, ε_{rep}	Measurement noise
H_0	Null hypothesis
H_A	Alternative hypothesis
α	Significance level
O_m	Observed counts
E_m	Expected counts
$\mathcal{Z}_{kl}$	Number of cells in a cell cycle phase
A_{ka} with $a = 1, ..., nR$	Measured number of mRNA molecules
$\mathbf{1} = (1, 1, ...)$	Vector of all entries equal to 1
y_{klm} with $l = 1, ..., nP$ and $m = 1, ..., nB$	Measured mRNA freqencies in cell cycle phase l and bin m
γ_{kl}	Simulated mRNA freqencies in cell cycle phase l
γ_{klm} with $l = 1, ..., nP$ and $m = 1, ..., nB$	Simulated mRNA freqencies in cell cycle phase l and bin m
$L_{pre}(y_k\|\theta^{pre})$	Likelihood function of the pre-processing
$L(y_k\|\theta)$	Likelihood function
$L(\boldsymbol{y}\|\theta)$	Likelihood function of the combined data
$L_{mRNA}(y_k\|\theta^k)$	Likelihood function of the mRNA optimization step
$L_{protein}(\boldsymbol{y}\|\theta)$	Likelihood function of the protein optimization step
$L'_{protein}(\boldsymbol{y}\|\theta)$	Likelihood function of the protein optimization step regarding $p_{Gauss}(\theta)$
$L''_{protein}(\boldsymbol{y}\|\theta)$	Likelihood function of the protein optimization step regarding $p_{Gauss}(\theta)$ and $p_{mRNA}(\theta^r)$
$\ell(\boldsymbol{y}\|\theta)$	Log-likelihood function of the combined data
$\ell_{mRNA}(y_k\|\theta^k)$	Log-likelihood function of the mRNA optimization step
$\ell_{protein}(\boldsymbol{y}\|\theta)$	log-likelihood function of the protein optimization step
$\ell'_{protein}(\boldsymbol{y}\|\theta)$	log-transform of $L'_{protein}(y_k\|\theta)$
$\ell''_{protein}(\boldsymbol{y}\|\theta)$	log-transform of $L''_{protein}(y_k\|\theta)$
$\ell_\eta(\boldsymbol{y}\|\theta^\eta)$	η specific log-likelihood function associated with L_1 regularization
$\ell_\eta^{protein}(\boldsymbol{y}\|\theta^{p,\eta})$	η specific log-likelihood function associated with L_1 regularization for protein parameters
$\ell_\eta^{mRNA}(\boldsymbol{y}\|\tilde{r}^\eta)$	η specific log-likelihood function associated with L_1 regularization for mRNA fold changes
$\ell_{L_\rho}(\boldsymbol{y}\|\theta)$	Penalized log-likelihood function
$\ell_{L_1}(\boldsymbol{y}\|\theta)$	L_1 regularized log-likelihood function
$\ell_{L_1}^{protein}(\boldsymbol{y}\|\theta^p)$	L_1 regularized log-likelihood function for protein parameters
$\ell_{L_1}^{mRNA}(\boldsymbol{y}\|\tilde{r})$	L_1 regularized log-likelihood function for mRNA fold changes
$PL(\boldsymbol{y}\|\theta_{\underline{\iota}})$	Profile likelihood
$PL_{mRNA}(y_k\|\theta^k_{\underline{\iota}})$	Profile likelihood of the mRNA optimization step
$PL_{protein}(\boldsymbol{y}\|\theta_{\underline{\iota}})$	Profile likelihood of the protein optimization step
$PL_\eta(\boldsymbol{y}\|\theta^\eta_{\underline{\iota}})$	Profile likelihood after L_1 regularization

$PL_{\eta}^{protein}(\boldsymbol{y}\|\theta_{\underline{\iota}}^{p,\eta})$	Profile likelihood after L_1 regularization for protein parameters
$PL_{\eta}^{mRNA}(\boldsymbol{y}\|\tilde{r}_{\underline{\iota}}^{\eta})$	Profile likelihood after L_1 regularization for mRNA fold changes
$D(\eta)$	Likelihood ratio associated with L_1 regularization
$D_{protein}(\eta)$	Likelihood ratio associated with L_1 regularization for protein parameters
$D_{mRNA}(\eta)$	Likelihood ratio associated with L_1 regularization for mRNA fold changes
η	Regularization strength
$\hat{\eta}$	Optimal regularization strength

B. Abbreviations

Abbreviation	Description
smFISH	Single molecule RNA *in situ* hybridization
FISSEQ	Fluorescent *in situ* sequencing
CDK	Cyclin dependent kinases
CKI	Cyclin dependent kinase inhibitors
G1	G1 phase
S	S phase
G2	G2 phase
P/M	Pro-/metaphase
Ana	Anaphase
T/C	Telophase/cytokinesis
SD	Standard deviation
SEM	Standard error of the mean
CKE	Chapman-Kolmogorov equation
CME	Chemical Master equation
SSA	Stochastic simulation algorithm
ODE	Ordinary differential equation
RRE	Reaction rate equation
PDF	Probability density function
CDF	Cumulative distribution function
MLE	Maximum likelihood estimate
CLT	Central limit theorem
iid	Identically and independently distributed
IQR	Interquartile range
LHS	Left-hand side
RHS	Right-hand side
LRHS	Left- and right-hand side

C. Definitions

C.1 Definition of time points t_{ln}

Time points t_{ln} in cell cycle phase l with interval lengths taken from [52] and $\Delta t = 0.1$ are given by

$$\begin{aligned} t_{1n} &\in [0.0, 27.3), \\ t_{2n} &\in [27.3, 61.0), \\ t_{3n} &\in [61.0, 81.5), \\ t_{4n} &\in [81.5, 102.9), \\ t_{5n} &\in [102.9, 112.7), \\ t_{6n} &\in [112.7, 123.1), \\ t_{7n} &\in [123.1, 127.1). \end{aligned} \tag{C.1}$$

C.2 Definition of the number of cells $\mathcal{Z}_{kl}$

The number of cells $\mathcal{Z}_{kl}$ in experimental data y_k and cell cycle phase l taken from [52] are given by

$$\begin{aligned} \mathcal{Z}_{1l} &= \{201, 276, 169, 143, 69, 69, 30\}, \\ \mathcal{Z}_{3l} &= \{236, 235, 121, 168, 74, 67, 27\}, \\ \mathcal{Z}_{5l} &= \{168, 259, 179, 145, 71, 88, 29\}. \end{aligned} \tag{C.2}$$

C.3 Definition of the number of mRNA molecules A_{ka}

The number of mRNA molecules A_{ka} in experimental data y_k taken from [52] are given by

$$\begin{aligned} A_{1a} &\in [0, 66], \\ A_{3a} &\in [0, 71], \\ A_{5a} &\in [0, 17]. \end{aligned} \tag{C.3}$$

D. Calculations

D.1 Calculation of pre-processed data by using blotIt2

We start from equation

$$y_{k,rep}(t_n) = f_{k,rep}(\theta^{pre}, t_n) + \varepsilon_{rep} \tag{D.1}$$

where $y_{k,rep}(t_n)$ is the measured value of observable y_k in replicate rep at time t_n, $f_{k,rep}(\theta^{pre}, t_n) = \tilde{y}_k(t_n)/s_{rep}$ is the response of the biological system at time t_n with time course parameter $\tilde{y}_k(t_n)$ and scaling parameter s_{rep}, and ε_{rep} is the measurement error ε_{rep} with $\varepsilon_{rep} \sim \mathcal{N}(0, \sigma_{rel}^2 f_{k,rep}(\theta^{pre}, t_n)^2)$ regarding a constant relative error $\sigma_{rel}^2 f_{k,rep}(\theta^{pre}, t_n)^2$. Using the maximum likelihood approach, the likelihood function

$$L_{pre}(y_k|\theta^{pre}) = \prod_{rep}^{nRep} \prod_{n}^{nT} \frac{1}{\sqrt{2\pi\sigma_{rel}^2 f_{k,rep}(\theta^{pre}, t_n)^2}} e^{-\frac{\left(y_{k,rep(t_n)} - f_{k,rep}(\theta^{pre}, t_n)\right)^2}{2\sigma_{rel}^2 f_{k,rep}(\theta^{pre}, t_n)^2}} \tag{D.2}$$

is optimized for parameters $\theta^{pre} = (\tilde{y}_k(t_{n=1}), ..., \tilde{y}_k(t_{n=nT}), s_{rep=1}, ..., s_{rep=nRep})$ with $nRep$ the number of replicates and nT the number of time points. For numerical reasons the negative of the log-likelihood function is minimized and parameters are log-transformed to ensure positive parameter estimates. If measured data are multiplicative log-normally distributed, $y_{k,rep}(t_n)$ and $f_{k,rep}(\theta^{pre}, t_n)$ can be log-transformed.

D.2 Calculation of the time evolution of the expected state

We start from Equation (4.21)

$$\langle \boldsymbol{X}(t) \rangle = \sum_{\boldsymbol{x}} \boldsymbol{x} P(\boldsymbol{x}, t|\boldsymbol{x}_0, t_0)$$

and calculate the time derivative of it

$$d_t \langle \boldsymbol{X}(t) \rangle = \sum_{\boldsymbol{x}} \boldsymbol{x}\, d_t P(\boldsymbol{x}, t|\boldsymbol{x}_0, t_0).$$

Now we insert Equation (4.13)

$$\begin{aligned} d_t \langle \boldsymbol{X}(t) \rangle &= \sum_{\boldsymbol{x}} \boldsymbol{x} \sum_{j=1}^{M} [a_j(\boldsymbol{x} - \boldsymbol{\nu}_j) P(\boldsymbol{x} - \boldsymbol{\nu}_j, t|\boldsymbol{x}_0, t_0) - a_j(\boldsymbol{x}) P(\boldsymbol{x}, t|\boldsymbol{x}_0, t_0)] \\ &= \sum_{j=1}^{M} \left[\sum_{\boldsymbol{x}} \boldsymbol{x} a_j(\boldsymbol{x} - \boldsymbol{\nu}_j) P(\boldsymbol{x} - \boldsymbol{\nu}_j, t|\boldsymbol{x}_0, t_0) - \sum_{\boldsymbol{x}} \boldsymbol{x} a_j(\boldsymbol{x}) P(\boldsymbol{x}, t|\boldsymbol{x}_0, t_0) \right]. \end{aligned}$$

With $\sum_{\boldsymbol{x}} \boldsymbol{x} a_j(\boldsymbol{x} - \boldsymbol{\nu}_j) P(\boldsymbol{x} - \boldsymbol{\nu}_j, t|\boldsymbol{x}_0, t_0)$ being equivalent to $\sum_{\boldsymbol{x}} (\boldsymbol{x} + \boldsymbol{\nu}_j) a_j(\boldsymbol{x}) P(\boldsymbol{x}, t|\boldsymbol{x}_0, t_0)$ we get

$$\begin{aligned} &= \sum_{j=1}^{M} \left[\sum_{\boldsymbol{x}} (\boldsymbol{x} + \boldsymbol{\nu}_j) a_j(\boldsymbol{x}) P(\boldsymbol{x}, t|\boldsymbol{x}_0, t_0) - \sum_{\boldsymbol{x}} \boldsymbol{x} a_j(\boldsymbol{x}) P(\boldsymbol{x}, t|\boldsymbol{x}_0, t_0) \right] \\ &= \sum_{j=1}^{M} \left[\sum_{\boldsymbol{x}} \boldsymbol{x} a_j(\boldsymbol{x}) P(\boldsymbol{x}, t|\boldsymbol{x}_0, t_0) + \sum_{\boldsymbol{x}} \boldsymbol{\nu}_j a_j(\boldsymbol{x}) P(\boldsymbol{x}, t|\boldsymbol{x}_0, t_0) - \sum_{\boldsymbol{x}} \boldsymbol{x} a_j(\boldsymbol{x}) P(\boldsymbol{x}, t|\boldsymbol{x}_0, t_0) \right] \\ &= \sum_{j=1}^{M} \boldsymbol{\nu}_j \sum_{\boldsymbol{x}} a_j(\boldsymbol{x}) P(\boldsymbol{x}, t|\boldsymbol{x}_0, t_0). \end{aligned}$$

Finally, we make use of Equation (4.22) and receive Equation (4.23):

$$d_t \langle \boldsymbol{X}(t) \rangle = \sum_{j=1}^{M} \boldsymbol{\nu}_j \langle a_j(\boldsymbol{X}(t)) \rangle.$$

D.3 Calculation of the time until next reaction for time dependent stochastic rates

For illustration purposes, we look at the pure birth process of species S_1 which is known as Poisson process: $\varnothing \xrightarrow{c_1(t)} S_1$. The time to the first reaction τ of a Poisson process is an exponential random variable with probability density function

$$f_{Exp}(\tau) = c_1(\tau)e^{-\Lambda(\tau)} \tag{D.3}$$

and cumulative distribution function

$$F_{Exp}(\tau) = 1 - e^{-\Lambda(\tau)} \tag{D.4}$$

for a time dependent transition rate $c_1(t)$ where $\Lambda(\tau) = \int_0^\tau c_1(\tau')d\tau'$. The integral becomes

$$\begin{aligned}\Lambda(\tau) &= \int_0^{t_{s_{1,0}}} c_1(\tau')d\tau' + \int_{t_{s_{1,0}}}^{t_{s_{1,e}}} c_1(\tau')d\tau' + \int_{t_{s_{1,e}}}^{\tau} c_1(\tau')d\tau' \\ &= 2c_{1,low}\tau + (c_{1,high} - c_{1,low})t_{s_{1,e}} - c_{1,high}t_{s_{1,0}}\end{aligned} \tag{D.5}$$

for

$$c_1(t) = \begin{cases} c_{1,high}, & t_{s_{1,0}} < t < t_{s_{1,e}} \\ c_{1,low}, & \text{otherwise.} \end{cases} \tag{D.6}$$

This exponential distribution describes not only the time to the first reaction but also the time between any two reaction events. To calculate the time until next reaction τ from a number ξ_1 which is sampled from the uniform distribution in the unit interval by inversion, we set $F_{Exp}(\tau) = \xi_1$ and get

$$\tau = \frac{1}{c_{1,low}}\left(\ln\left(\frac{1}{\xi_1}\right) + (c_{1,high} - c_{1,low})t_{s_{1,e}} - c_{1,high}t_{s_{1,0}}\right). \tag{D.7}$$

D.4 Calculation of the stationary distribution of a birth-death process

We derive the stationary distribution from a simplified system $\varnothing \xrightarrow{c_1} S_1 \xrightarrow{c_2} \varnothing$ which includes the inflow (R_1) and the outflow (R_2) reaction of species S_1 and is characterized by state-change vectors $\nu_1 = +1$ and $\nu_2 = -1$ and propensity functions $a_1(x) = c_1$ and $a_2(x) = c_2x$. The CME (see Equation (4.13)) for this system becomes

$$\begin{aligned}d_tP(x,t|x_0,t_0) &= c_1P(x-1,t|x_0,t_0) + c_2(x+1)P(x+1,t|x_0,t_0) \\ &\quad - c_1P(x,t|x_0,t_0) - c_2xP(x,t|x_0,t_0).\end{aligned} \tag{D.8}$$

The CME tends in the limit $t \to \infty$ to its stationary distribution $P^\star(x,t)$ which is reached if the probability density function $P(x,t|x_0,t_0)$ (see Definition 4.2) does not change over time. We use the ansatz: $d_tP^\star(x,t) = 0$. The stationary distribution is no longer conditioned on the previous system state. What we get out of it is

$$\begin{aligned}&c_1P^\star(x-1,t) - c_2xP^\star(x,t) \\ &= c_1P^\star(x,t) - c_2(x+1)P^\star(x+1,t)\end{aligned} \tag{D.9}$$

and is named detailed balance. Since this balance equation holds true for all x, both sides are equal to a constant κ [134]. To determine the value of the constant κ, we look at the righthand side of the balnce equation (D.9)

$$c_1P^\star(x,t) - c_2(x+1)P^\star(x+1,t) = \kappa. \tag{D.10}$$

Now we calculate the sum over all x

$$c_1\sum_x P^\star(x,t) - c_2\sum_x (x+1)P^\star(x+1,t) = \sum_x \kappa, \tag{D.11}$$

make use of $P^\star(x,t)$ being normalized $\sum_x P^\star(x,t) = 1$ and $\sum_x(x+1)P^\star(x+1,t) = \sum_x(x)P^\star(x,t)$ to get

$$c_1 - c_2\sum_x xP^\star(x,t) = \sum_x \kappa. \tag{D.12}$$

Using $\sum_x xP^\star(x,t) = \langle x^\star \rangle$ and $\langle x^\star \rangle = \frac{c_1}{c_2}$ in steady state we end up with

$$\sum_x \kappa = 0. \tag{D.13}$$

This equality is only fulfilled if $\kappa = 0$.

To further determine the stationary distribution, we look at the lefthand side of the balance equation (D.9)

$$c_1 P^\star(x-1,t) - c_2 x P^\star(x,t) = 0 \tag{D.14}$$

which is equivalent to

$$c_1 P^\star(x-1,t) = c_2 x P^\star(x,t). \tag{D.15}$$

Here, we can see that the flow of probability between $x-1$ and x is equal, which holds true for any pair of adjacent states. Rearrangement leads to

$$\begin{aligned} P^\star(x,t) &= \frac{c_1}{c_2}\frac{1}{x} P^\star(x-1,t) \\ &= \left(\frac{c_1}{c_2}\right)^2 \frac{1}{x(x-1)} P^\star(x-2,t) \\ &= ... \\ &= \left(\frac{c_1}{c_2}\right)^x \frac{1}{x!} P^\star(0,t). \end{aligned} \tag{D.16}$$

Using again $\sum_x P^\star(x,t) = 1$ gives

$$\begin{aligned} P^\star(0,t) \sum_x \left(\frac{c_1}{c_2}\right)^x \frac{1}{x!} &= 1 \\ P^\star(0,t) e^{\frac{c_1}{c_2}} &= 1 \end{aligned} \tag{D.17}$$

which is only true if $P^\star(0,t) = e^{-\frac{c_1}{c_2}}$. Finally, the steady state distribution is a Poisson distribution

$$P^\star(x,t) = \frac{(\lambda^\star)^x}{x!} e^{-\lambda^\star} \tag{D.18}$$

with $\lambda^\star = \frac{c1}{c2}$. Depending on the choice of stochastic rate constants c_1 and c_2, the system can also converge to the extinction of the species. However, in this study we assume that a unique stationary distribution exists and do not take into account other cases.

D.5 Calculation of the analytical solution of the CME for a birth-death process with a Poisson initial distribution

We derive the analytical solution from a reduced system $\varnothing \xrightarrow{c_1(t)} S_1 \xrightarrow{c_2} \varnothing$ which includes the inflow (R_1) and the outflow (R_2) reaction of species S_1 and is characterized by state-change vectors $\nu_1 = +1$ and $\nu_2 = -1$ and the propensity functions $a_1(x) = c_1(t)$ and $a_2(x) = c_2 x$. The CME (see Equation (4.13)) for this system becomes

$$\begin{aligned} d_t P(x,t|x_0,t_0) = c_1(t)P(x-1,t|x_0,t_0) + c_2(x+1)P(x+1,t|x_0,t_0) \\ - c_1(t)P(x,t|x_0,t_0) - c_2 x P(x,t|x_0,t_0). \end{aligned} \tag{D.19}$$

We start from a Poisson initial distribution

$$P(x,0) = \mathcal{P}(x,\lambda_0) \tag{D.20}$$

to show that the solution of the CME is still a Poisson distribution

$$\begin{aligned} P(x,t|x_0,t_0) &= \mathcal{P}(x,\lambda(t)) \\ &= \frac{\lambda(t)^x}{x!} e^{-\lambda(t)} \end{aligned} \tag{D.21}$$

with a time dependent mean value which follows the RRE (see Equation 4.18) of the reduced system

$$\begin{aligned} d_t\lambda(t) &= c_1(t) - c_2\lambda(t), \\ \lambda(0) &= \lambda_0. \end{aligned} \tag{D.22}$$

The derivative of the Poisson distribution is

$$\begin{aligned} d_t\mathcal{P}(x,\lambda(t)) &= -d_t\lambda(t)\frac{\lambda(t)^x}{x!}e^{-\lambda(t)} + d_t\lambda(t)\frac{\lambda(t)^x - 1}{(x-1)!}e^{-\lambda(t)} \\ &= -d_t\lambda(t)\mathcal{P}(x,\lambda(t) + d_t\lambda(t)\mathcal{P}(x-1,\lambda(t). \end{aligned} \tag{D.23}$$

The first term gives

$$\begin{aligned} -d_t\lambda(t)\mathcal{P}(x,\lambda(t) &= -c_1(t)\mathcal{P}(x,\lambda(t) + c_2\lambda(t)\mathcal{P}(x,\lambda(t) \\ &= -c_1(t)\mathcal{P}(x,\lambda(t) + c_2(x+1)\mathcal{P}(x+1,\lambda(t) \end{aligned} \tag{D.24}$$

and the second term

$$\begin{aligned} d_t\lambda(t)\mathcal{P}(x-1,\lambda(t) &= c_1(t)\mathcal{P}(x-1,\lambda(t) - c_2\lambda(t)\mathcal{P}(x-1,\lambda(t) \\ &= c_1(t)\mathcal{P}(x-1,\lambda(t) - c_2x\mathcal{P}(x,\lambda(t). \end{aligned} \tag{D.25}$$

Finally Equation D.23 becomes

$$\begin{aligned} d_t\mathcal{P}(x,\lambda(t)) &= c_1(t)\mathcal{P}(x-1,\lambda(t) + c_2(x+1)\mathcal{P}(x+1,\lambda(t) \\ &\quad - c_1(t)\mathcal{P}(x,\lambda(t) - c_2x\mathcal{P}(x,\lambda(t), \end{aligned} \tag{D.26}$$

which is exactly Equation D.19 for $P(x,t|x_0,t_0) = \mathcal{P}(x,\lambda(t))$. This solution is derived form the general solution presented in [95].

D.6 Calculation of the gradient of the objective function in the mRNA optimization step

Calculating the gradient $d_{\theta^k}\ell_{mRNA}(y_k|\theta^k)$ means to compute derivatives of the negative log-likelihood function with respect to each parameter θ^k_ι to be estimated. The negative log-likelihood function of the mRNA optimization step is

$$\ell_{mRNA}(y_k|\theta^k) = \sum_{l=1}^{nP}\sum_{m=1}^{nB}\left[\gamma_{klm} + \log(y_{klm}!) - y_{klm}\log(\gamma_{klm})\right] \tag{D.27}$$

and the derivatives become

$$\begin{aligned} d_{\theta^k_\iota}\ell_{mRNA}(y_k|\theta^k) &= \sum_{l=1}^{nP}\sum_{m=1}^{nB}\left[d_{\theta^k_\iota}\big(\gamma_{klm}\big) + d_{\theta^k_\iota}\big(\log(y_{klm}!)\big)\right. \\ &\quad \left. - d_{\theta^k_\iota}\big(y_{klm}\log(\gamma_{klm})\big)\right] \\ &= \sum_{l=1}^{nP}\sum_{m=1}^{nB}\left[d_{\theta^k_\iota}\big(\gamma_{klm}\big) - \frac{y_{klm}}{\gamma_{klm}}d_{\theta^k_\iota}\big(\gamma_{klm}\big)\right] \\ &= \sum_{l=1}^{nP}\sum_{m=1}^{nB}\left[\gamma'_{klm} - \frac{y_{klm}}{\gamma_{klm}}\gamma'_{klm}\right], \end{aligned} \tag{D.28}$$

with

$$\gamma'_{kl} = d_{\theta^k_\iota}\big(\gamma_{kl}\big) = \frac{\mathcal{Z}_{kl}}{nT}\cdot\boldsymbol{\mathcal{P}}'\cdot\mathbf{1} \tag{D.29}$$

and

$$\begin{aligned} \boldsymbol{\mathcal{P}}' = d_{\theta^k_\iota}\boldsymbol{\mathcal{P}} &= \{d_{\theta^k_\iota}\mathcal{P}_{an}\} \\ &= \left\{d_{\theta^k_\iota}\left(\frac{Y_k(t_{ln},\theta^k)^{A_{ka}}}{A_{ka}!}e^{-Y_k(t_{ln},\theta^k)}\right)\right\} \\ &= \left\{\frac{Y_k(t_{ln},\theta^k)^{(\mathcal{A}(y_k))_a}}{A_{ka}!}e^{-Y_k(t_{ln},\theta^k)}\right. \\ &\quad \left.\cdot\left(\frac{A_{ka}}{Y_k(t_{ln},\theta^k)} - 1\right)d_{\theta^k_\iota}Y_k(t_{ln},\theta^k)\right\}. \end{aligned} \tag{D.30}$$

Since $Y_k(t,\theta^k) = C_i(t,\theta^k)$ for $k = \{1,3,5\}$ (see Table 5.1), $Y_k(t_{ln},\theta^k) = C_i(t_{ln},\theta^k)$ and $d_{\theta^k_\iota}Y_k(t_{ln},\theta^k) = d_{\theta^k_\iota}C_i(t_{ln},\theta^k)$. Thus, sensitivities of observables are equal to model sensitivities.

D.7 Calculation of the gradient of the objective function in the protein optimization step

Calculating the gradient $d_\theta \ell''_{protein}(\boldsymbol{y}|\theta)$ means to compute derivatives of the negative log-likelihood function with respect to each parameter θ_ι to be estimated. The negative log-likelihood function of the protein optimization step in consideration of prior knowledge (Gaussisan and mRNA prior) is

$$\begin{aligned}\ell''_{protein}(\boldsymbol{y}|\theta) = &\sum_{k=\{2,4,6\}}\sum_{n=1}^{nT}\frac{1}{2}\left[\left(\frac{y_{kn}-Y_k(t_n,\theta)}{\sigma_{kn}}\right)^2+\log(2\pi\sigma_{kn}^2)\right]\\ &+\sum_{\iota=1}^{nPar}\frac{1}{2}\left[\left(\frac{\theta^*_\iota-\theta_\iota}{\sigma_{\theta_\iota}}\right)^2+\log(2\pi\sigma_{\theta_\iota}^2)\right]\\ &+\sum_{\iota=1}^{nPar}\frac{1}{2}\left[\left(\frac{\theta^{r*}_\iota-\theta^r_\iota}{\sigma_{\theta^r_\iota}}\right)^2+\log(2\pi\sigma_{\theta^r_\iota}^2)\right]\end{aligned} \tag{D.31}$$

and the derivatives becomes

$$\begin{aligned}d_{\theta_\iota}\ell''_{protein}(\boldsymbol{y}|\theta) = &\sum_{k=\{2,4,6\}}\sum_{n=1}^{nT}\frac{1}{2\sigma_{kn}^2}\left[d_{\theta_\iota}\left(y_{kn}-Y_k(t_n,\theta)\right)^2\right]\\ &+\frac{1}{2\sigma_{\theta_\iota}}\left[d_{\theta_\iota}\left(\theta^*_\iota-\theta_\iota\right)^2\right]+\frac{1}{2\sigma_{\theta^r_\iota}}\left[d_{\theta^r_\iota}\left(\theta^{r*}_\iota-\theta^r_\iota\right)^2\right]\\ = &\sum_{k=\{2,4,6\}}\sum_{n=1}^{nT}-\frac{1}{\sigma_{kn}^2}\left(y_{kn}-Y_k(t_n,\theta)\right)d_{\theta_\iota}Y_k(t_n,\theta)\\ &-\frac{1}{\sigma_{\theta_\iota}}\left(\theta^*_\iota-\theta_\iota\right)-\frac{1}{\sigma_{\theta^r_\iota}}\left(\theta^{r*}_\iota-\theta^r_\iota\right)\end{aligned} \tag{D.32}$$

where the last term only exists if $\theta_\iota = \theta^r_\iota$. Sensitivities of observables are

$$d_{\theta_\iota}\boldsymbol{Y}(t,\theta) = d_{\boldsymbol{C}(t,\theta)}\boldsymbol{g}(\boldsymbol{C}(t,\theta),\theta)\cdot d_{\theta\iota}\boldsymbol{C}(t,\theta)+d_{\theta_\iota}\boldsymbol{g}(\boldsymbol{C}(t,\theta),\theta). \tag{D.33}$$

Regarding the definition of the observation function $\boldsymbol{g}(\boldsymbol{C}(t,\theta),\theta)$ (see Table 5.1), $d_{\boldsymbol{C}(t,\theta)}\boldsymbol{g}(\boldsymbol{C}(t,\theta),\theta)=1$ and $d_{\theta_\iota}\boldsymbol{g}(\boldsymbol{C}(t,\theta),\theta)=0$ for all derivatives. Hence, $d_{\theta_\iota}\boldsymbol{Y}(t,\theta)=d_{\theta_\iota}\boldsymbol{C}(t,\theta)$ meaning that sensitivities of observables are equal to model sensitivities.

D.8 Calculation of time dependent mRNA production rates by using Fermi-Dirac distributions

We use Fermi-Dirac distributions to realize continuous transitions between high and low transcription rates. Species S_1 has only one high transcription region and the time dependent transcription rate becomes

$$k_1(t) \quad = k_{1,low}(1-FD)+k_{1,high}FD \tag{D.34}$$

with

$$FD = f_{FD}(t,t_{s_{1,e}},\beta)f_{1-FD}(t,t_{s_{1,0}},\beta) \tag{D.35}$$

where

$$f_{FD}(t,\vartheta,\beta) = \frac{1}{e^{(t-\vartheta)/\beta}+1} \tag{D.36}$$

and

$$\begin{aligned}f_{1-FD}(t,\vartheta,\beta) &= f_{FD}(t,\vartheta,\beta)\\ &= \frac{e^{(t-\vartheta)/\beta}}{e^{(t-\vartheta)/\beta}+1}.\end{aligned} \tag{D.37}$$

Parameter ϑ describes the transition point while β determines how fast the transition is. Smaller values result in a faster transition. We set $\beta = 0.5$ in the mRNA optimization step and $\beta = 0.6$ in the protein optimization step. Smaller values were not possible caused by integration errors.

Species S_3 and S_5 have a second high transcription region. The time dependent transcription rate for species S_3 becomes

$$\begin{aligned} k_5(t) = k_{5,low}(1 - (FD_{1st} + FD_{2nd})) \\ + k_{5,high}(FD_{1st} + FD_{2nd}) \end{aligned} \tag{D.38}$$

with

$$FD_{1st} = f_{FD}(t, t_{s_{2,e,1st}}, \beta) f_{1-FD}(t, t_{s_{2,0,1st}}, \beta) \tag{D.39}$$

and

$$FD_{2nd} = f_{FD}(t, t_{s_{2,e,2nd}}, \beta) f_{1-FD}(t, t_{s_{2,0,2nd}}, \beta). \tag{D.40}$$

The time dependent transcription rate for species S_5 is equivalent.

D.9 Calculation of p-values in a χ^2 goodness of fit test

We use the χ^2 goodness of fit test to evaluate if an observed distribution (data) is represented by a known probability density function (model), e.g. Poisson distribution [61]. Starting from a null and an alternative hypothesis

H_0: There is no significant difference between model and data
H_A: There is a significant difference between model and data,

we calculate the test statistic

$$\chi_t^2 = \sum_{m=1}^{nB} \frac{(O_m - E_m)^2}{E_m} \tag{D.41}$$

where nB is the number of bins, O_m are observed and E_m expected counts. Expected counts represent the product of the probability given by the model and the total number of data points. The test statistic is for sufficiently large observations O_m approximately $\chi^2(df)$ distributed. The degree of freedom df is determined by $df = nB - nPar - 1$, with $nPar$ the number of estimated parameters. Dependent on the degree of freedom df and the significance level α, we calculate a critical value χ_c^2. This value is the $(1-\alpha)$ quantile of the $\chi^2(df)$ distribution. We reject the null hypothesis H_0 if $\chi_t^2 > \chi_c^2$ and retain H_0 otherwise. Intuitively, χ_t^2 becomes large if the difference between model and data is large.

Instead of comparing the test statistic χ_t^2 with the critical value χ_c^2, it is more common to compare the significance level α with p-values. Assuming the null hypothesis H_0 is true, p-values indicate how likely our data sample is. p-values are calculated by

$$p\text{-value} = 1 - F_{\chi^2}(\chi_t^2) \tag{D.42}$$

where $F_{\chi^2}(\chi_t^2)$ is the cumulative distribution function of the $\chi^2(df)$ distribution. We reject the null hypothesis H_0 if p-value $< \alpha$ and retain H_0 otherwise.

D.10 Latin hypercube sampling versus Gaussian sampling

A multi-start optimization requires an efficient sampling method for guessing initial values for each optimization run which has to represent the whole parameter space. In a pure random sampling method, a random number for each parameter is drawn from a probability density function. We use a normal distribution where mean and variance have to be specified. We call a random sampling which is based on the normal distribution "Gaussian sampling". Further, we use the same variance for every parameter. Disadvantages of this sampling method are the possibility to draw nearby parameter values in successive samples as well as the under-representation of values which lie at the distribution tails. The latter can be reduced by correspondingly large variances.

In a latin hypercube sampling, we make use of the "sampling history" and therefore prevent nearby parameter values [77, 113, 135]. At first, upper and lower limits of each parameter and the number of samples N_{LHS} are chosen. Using these numbers an hypercube is build with N_{LHS} equal-sized segments on each parameter axis. For illustration purposes imagine a two-dimensional parameter space with N_{LHS}^2 equal boxes. A box is selected in the first row and a random number is drawn within this box from an uniform distribution which giving the first sample. A different box is selected in the second row and the second random sample is drawn. The change between selecting a box and drawing a random numbers goes on until N_{LHS} samples are drawn and each segment for each parameter is selected only ones. An disadvantage of this method is that we need at least some prior knowledge about the expected range of each parameter.

E. Figures

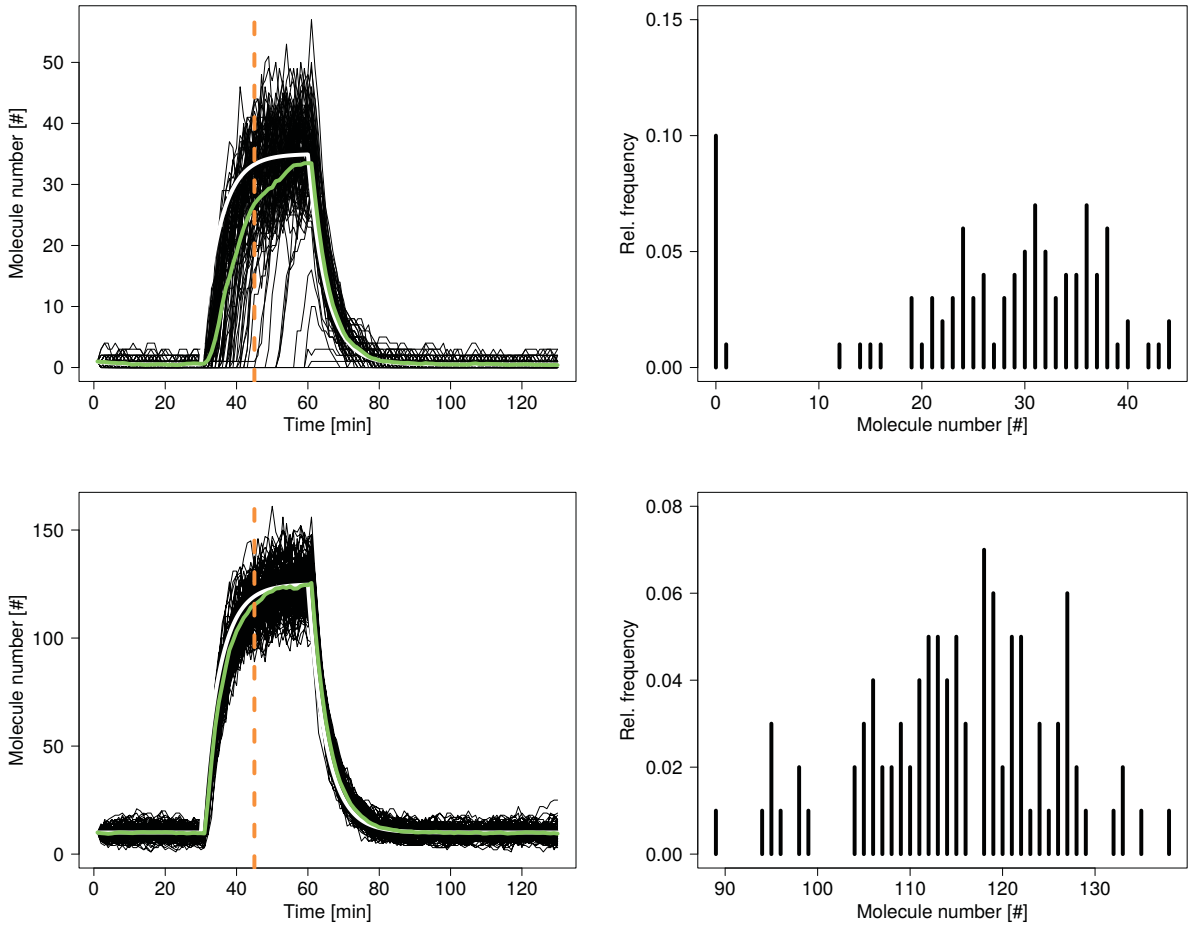

Figure E.1: Effect of high and low molecule numbers. In this figure, 100 realizations (solid black lines) of production and degradation of species S_1, $\varnothing \xrightarrow{c_1(t)} S_1 \xrightarrow{c_2} \varnothing$, for two different parameter settings are shown to visualize the effect of low and high molecule numbers. The production rate is time dependent $c_1(t)$ as defined in Table 4.2. The upper row presents effects of low molecule numbers. The transition from the low to the high transcription region occurs later for some realizations and not at all for others. Thus, the relative frequency of zero molecules becomes high in the distribution (orange dashed line). Even if we consider unimolecular reactions where the time evolution of the expected state is equal to the RRE, the mean value $\overline{X^{SSA}(t)}$ over all realizations (light green solid line) clearly differ from the solution of the corresponding RRE (white solid line). Both effects are removed if parameters are changed to get higher molecule numbers as shown in the lower row.

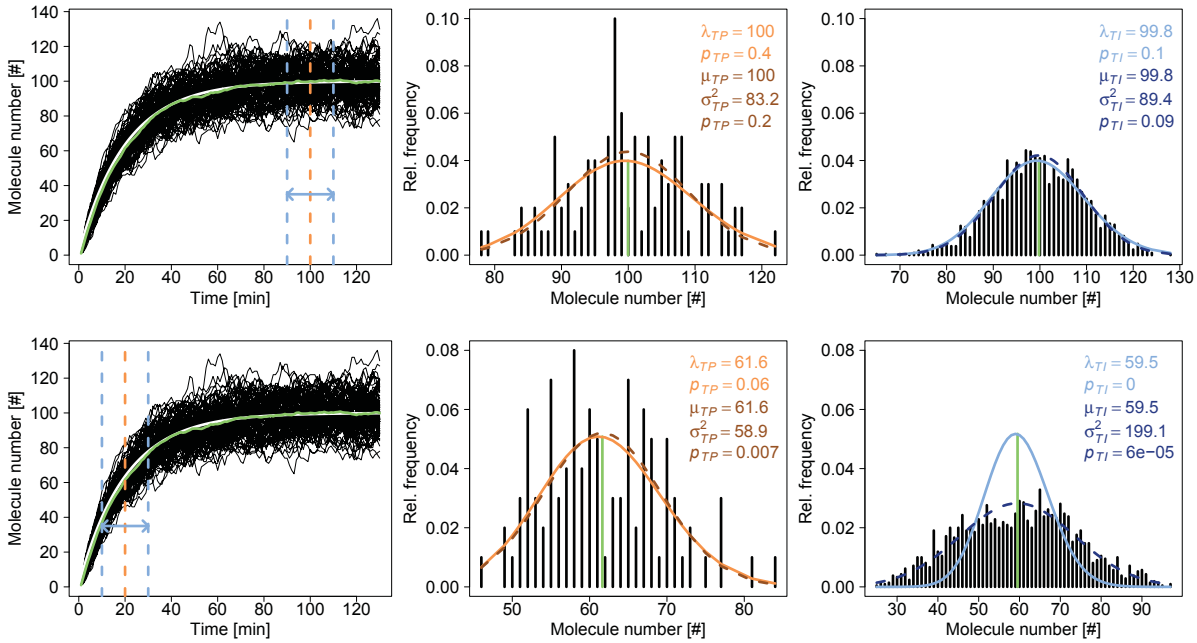

Figure E.2: Time point and time interval distributions for a constant mRNA production rate. In this figure, 100 realizations (solid black lines) of production and degradation of species S_1, $\varnothing \xrightarrow{c_1} S_1 \xrightarrow{c_2} \varnothing$, are shown to compare the average distribution over a time interval (TI, light blue) with the distribution at any time point within the time interval (TP, orange). The stochastic rate is a constant c_1 and the stationary distribution $P^\star(x,t)$ of this birth-death process is a Poisson distribution $\mathcal{P}(\lambda^\star)$ with $\lambda^\star = c1/c2$ (see Appendix D.4 for calculation details). The upper row shows that the average distribution over a time interval equals the Poisson distribution at any time point within the time interval in steady state, $\mathcal{P}_{TP}(\lambda^\star) = \mathcal{P}_{TI}(\lambda^\star)$. Estimated mean values, λ_{TP} and λ_{TI}, show small differences resulting from the simulation itself. In contrast, the average distribution over a time interval outside the steady state region is not well approximated by a Poisson distribution but the distribution at any time point within the time interval is Poisson distributed (lower row). Even if not significant, the average distribution is better described by a normal distribution (dark blue). In addition, the normal distribution and the Poisson distribution (dark orange and orange, middle column) are approximately the same for time point distributions resulting from a large avarage number of molecules ($\lambda \approx \mu \approx \sigma^2$). If the stochastic rate becomes time dependent $c_1(t)$ and the underlying function has a limit meaning that $\lim_{t\to\infty} c_1(t)$ exists, the behavior is the same as in the time independent case. An example is a sigmoid function $c_1(t) = c_1 \tanh(t)$. We used a χ^2 goodness of fit test to calculate p-values (see Appendix D.9 for calculation details). Large p-values compared to a significance level of $\alpha = 0.05$ indicate distributions reasonably represented by a Poisson or normal distribution. Lines between relative frequencies of Poisson distributions are for visualization only. All given values are rounded to the first decimal different from zero. The mean value of all realizations $\overline{X^{SSA}(t)}$ is represented by a solid light green line and the solution of the corresponding RRE by a white solid line.

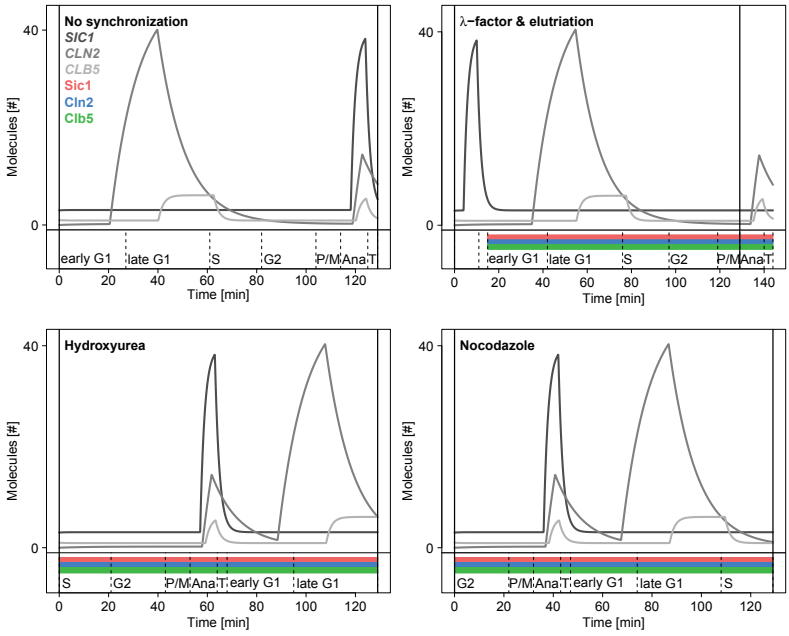

Figure E.3: Initial setting of the protein optimization step. In this figure, we represent how timing in the simulation of Equation 5.1 is changed to optimally estimate parameters from Western blot data (see Figure 8.3 and 8.2) in the protein optimization step. At the beginning of the optimization, we can only use results of the mRNA optimization step (no synchronization) and knowledge about synchronization methods why we plot mRNA time courses (gray lines) with the cell division time and cell cycle phases taken from [52]. Positioning of protein data is shown by colored bars. Time shifted simulations always cover one cell cycle passage. Elutriated and α-factor synchronized cells start in early G1 phase after release with a maximum Sic1 level. Furthermore, Sic1 predominantly decrease over time. To estimate the production rate of Sic1, we start simulation in anaphase, run over a correspondingly extended cell division time and shifted only the high transcription region of Sic1 to the first occurrence of ana- and T/C phase. Additionally, we fixed all initial values to zero. Hydroxyurea and nocodazole synchronized cells start in S and G2 phase after release, respectively. We start simulation in these phases and run over one cell division time, so that we consider parts of two successive cell cycles. Here, we estimated initial values. Finally, timing of high transcription regions and cell division times are shortened or enlarged during the optimization process. We distinguished between seven cell cycle phases: early G1, late G1, S, G2, pro-/metaphase (P/M), anaphase (Ana) and telophase/cytokinesis (T/C).

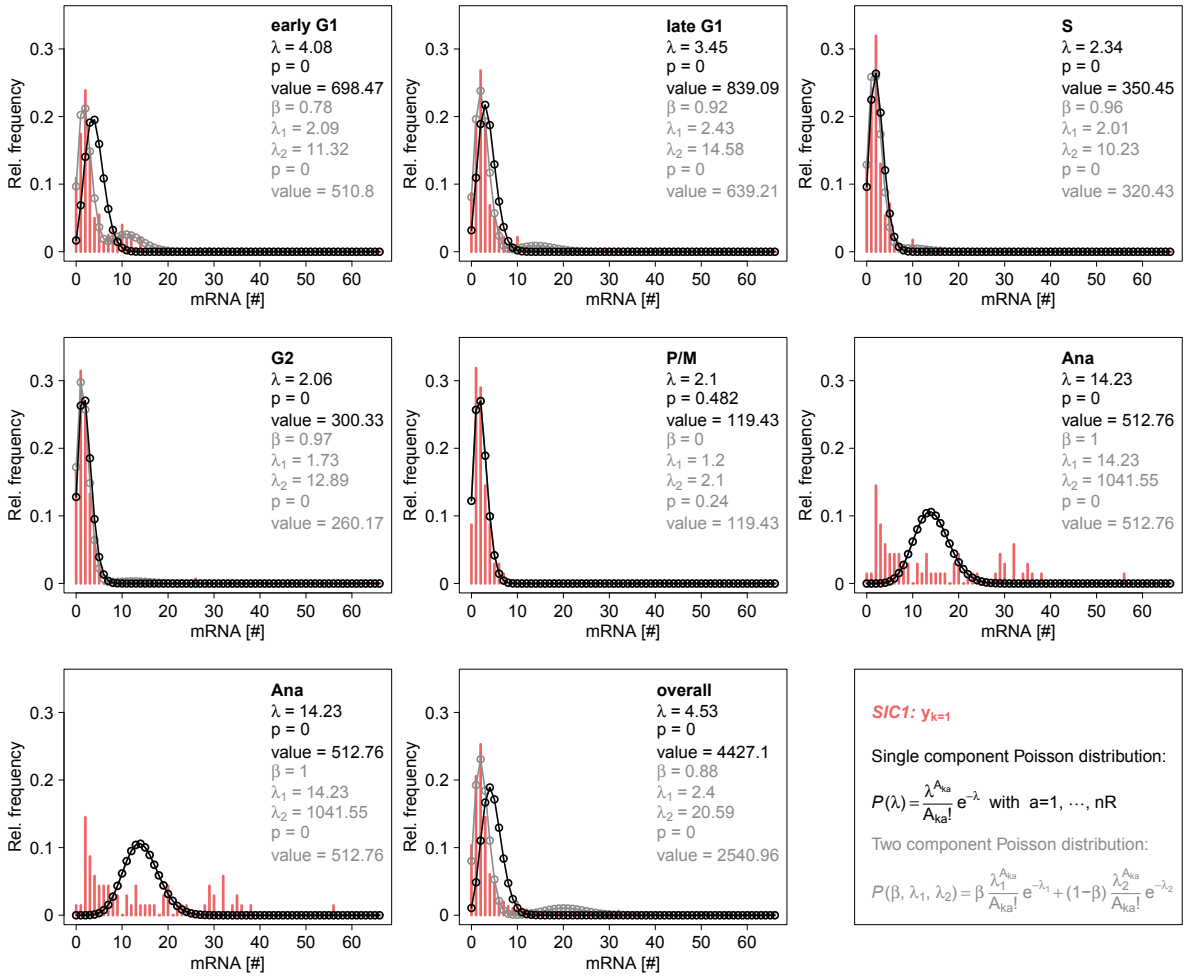

Figure E.4: Single and two component Poisson distributions fitted to smFISH data of mRNA species *SIC1*. In this figure, phase-resolved mRNA distributions of mRNA species *SIC1* are shown. For reasons of comparability, we plotted relative frequencies against the number of mRNA molecules. We distinguished between seven cell cycle phases: early G1, late G1, S, G2, pro-/metaphase (P/M), anaphase (Ana) and telophase/cytokinesis (T/C). Additionally, we plotted the mRNA distribution over the whole cell cycle. Open circles represent single (black) and two (gray) component Poisson distributions which are calculated from estimated parameters indicated in each plot. Connecting lines are for visualization only. Fitted probability density functions are given in the bottom right diagram where A_{ka} is the number of mRNA molecules dependent on experimental data y_k and running from $a = 1, ..., nR$ (see Equations C.3). p-values refer to the χ^2 goodness of fit test (see Appendix D.9 for calculation details). mRNA distribution of P/M is the only distribution which is not significantly different from both probability density functions. Values correspond to negative log-likelihood values.

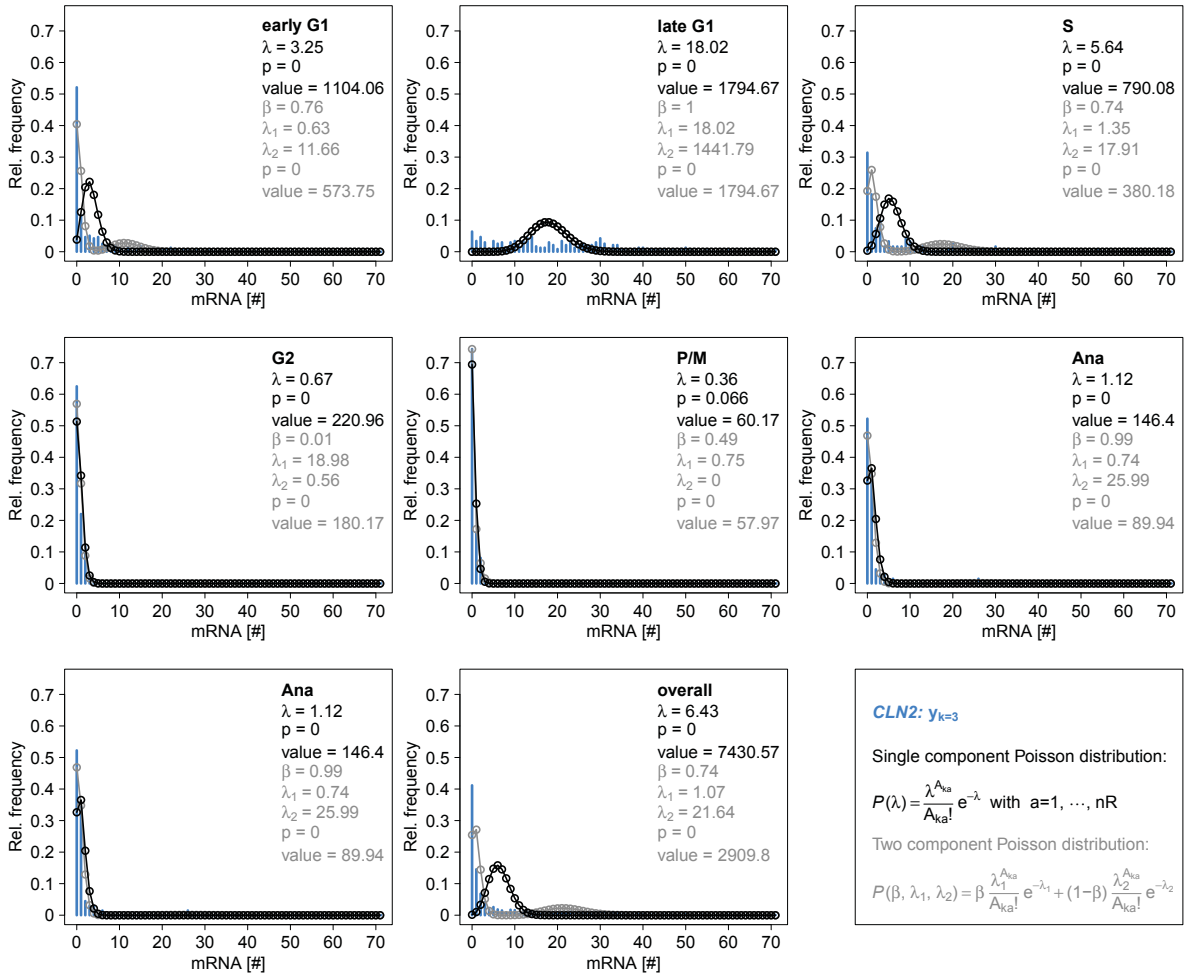

Figure E.5: Single and two component Poisson distributions fitted to smFISH data of mRNA species *CLN2*. In this figure, phase-resolved mRNA distributions of mRNA species *CLN2* are shown. For reasons of comparability, we plotted relative frequencies against the number of mRNA molecules. We distinguished between seven cell cycle phases: early G1, late G1, S, G2, pro-/metaphase (P/M), anaphase (Ana) and telophase/cytokinesis (T/C). Additionally, we plotted the mRNA distribution over the whole cell cycle. Open circles represent single (black) and two (gray) component Poisson distributions which are calculated from estimated parameters indicated in each plot. Connecting lines are for visualization only. Fitted probability density functions are given in the bottom right diagram where A_{ka} is the number of mRNA molecules dependent on experimental data y_k and running from $a = 1, ..., nR$ (see Equations C.3). p-values refer to the χ^2 goodness of fit test (see Appendix D.9 for calculation details). mRNA distribution of P/M is the only distribution which is not significantly different from both probability density functions. Values correspond to negative log-likelihood values.

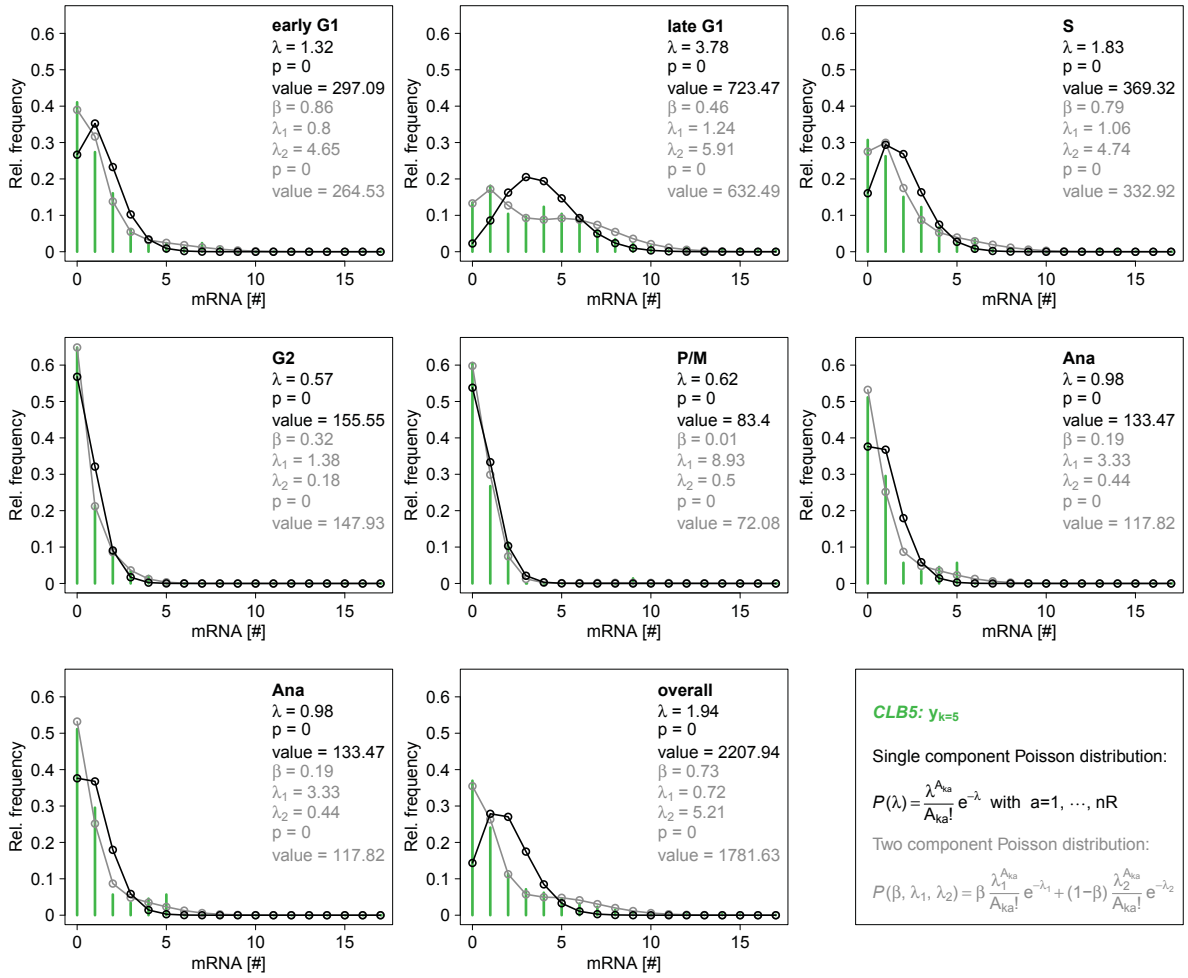

Figure E.6: Single and two component Poisson distributions fitted to smFISH data of mRNA species *CLB5*. In this figure, phase-resolved mRNA distributions of mRNA species *SIC1* are shown. For reasons of comparability, we plotted relative frequencies against the number of mRNA molecules. We distinguished between seven cell cycle phases: early G1, late G1, S, G2, pro-/metaphase (P/M), anaphase (Ana) and telophase/cytokinesis (T/C). Additionally, we plotted the mRNA distribution over the whole cell cycle. Open circles represent single (black) and two (gray) component Poisson distributions which are calculated from estimated parameters indicated in each plot. Connecting lines are for visualization only. Fitted probability density functions are given in the bottom right diagram where A_{ka} is the number of mRNA molecules dependent on experimental data y_k and running from $a = 1, ..., nR$ (see Equations C.3). p-values refer to the χ^2 goodness of fit test (see Appendix D.9 for calculation details). There is no mRNA distribution which is not significantly different from both probability density functions. Values correspond to negative log-likelihood values.

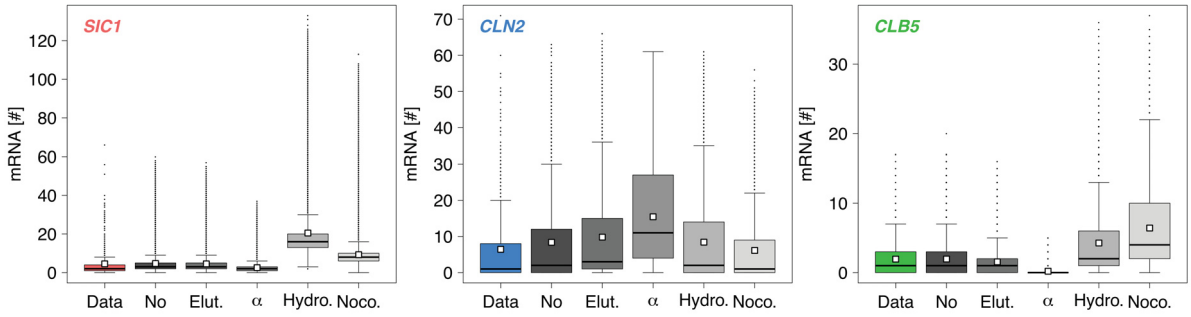

Figure E.7: Measured and simulated phase-resolved mRNA distributions over the whole cell cycle. In this figure, we show results of 2000 stochastic simulations for mRNA species *SIC1* (red), *CLN2* (blue) and *CLB5* (green) and synchronization specific parameter sets (grayscales), named "No" ("No synchronization"), "Elut." ("Elutriation"), "α" ("α-factor"), "Hydro." ("Hydroxyurea") and "Noco." ("Nocodazole"). smFISH data, named "Data", include measurements of more than 900 cells per mRNA species. We used integer valued phase lengths in the simulation (see Table 8.1) and simulated 129 time points ($\Delta t = 1$). "No synchronization" corresponds to simulations with parameters estimated from smFISH data in the mRNA optimization step. Other parameters are re-estimated from Western blot data in the protein optimization step. Distributions are represented as boxplots. First ($Q1$) and third quartile ($Q3$) form boxes with median (black lines) and mean (white squares). Whiskers range from $Q1 - 1.5 \cdot IQR$ to $Q3 + 1.5 \cdot IQR$ with an interquartile range of $IQR = Q3 - Q1$. Outliers are marked by black dots.

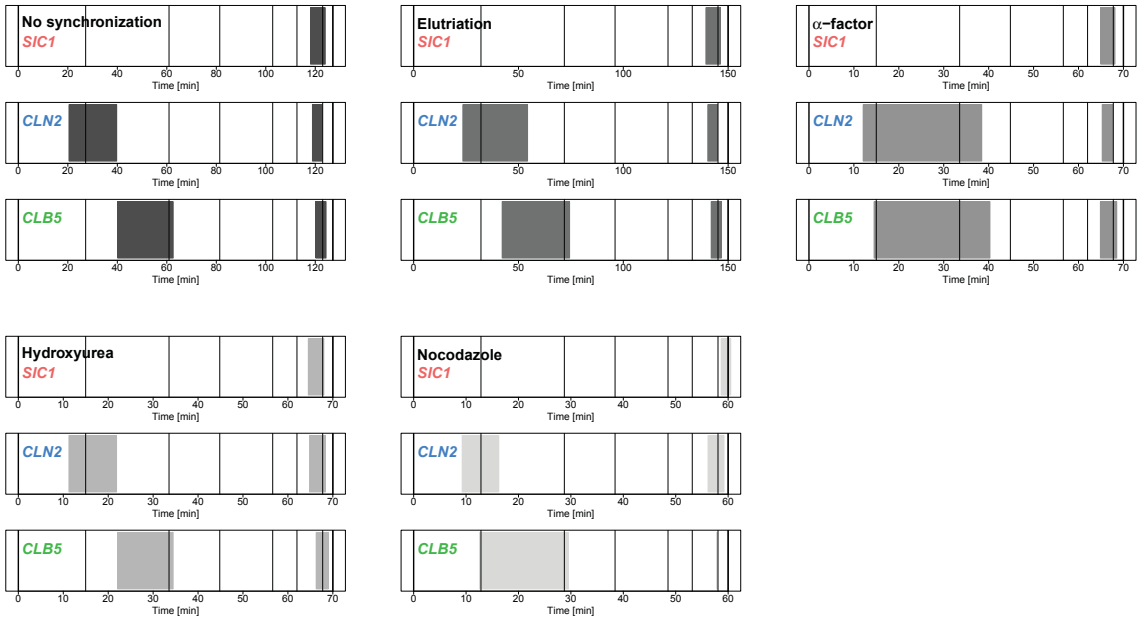

Figure E.8: Positioning of high transcription regions. In this figure, we illustrate positions of high transcription regions (grayscales) estimated for unsynchronized and synchronized cells. Timing parameters in "No synchronization" are estimated from smFISH data in the mRNA optimization step whereas other timing parameters are re-estimated from Western blot data in the protein optimization step. Vertical lines mark cell cycle phases relative to cell cycle phases determined for unsynchronized cells. The first cell cycle phase corresponds to the early G1 phase.

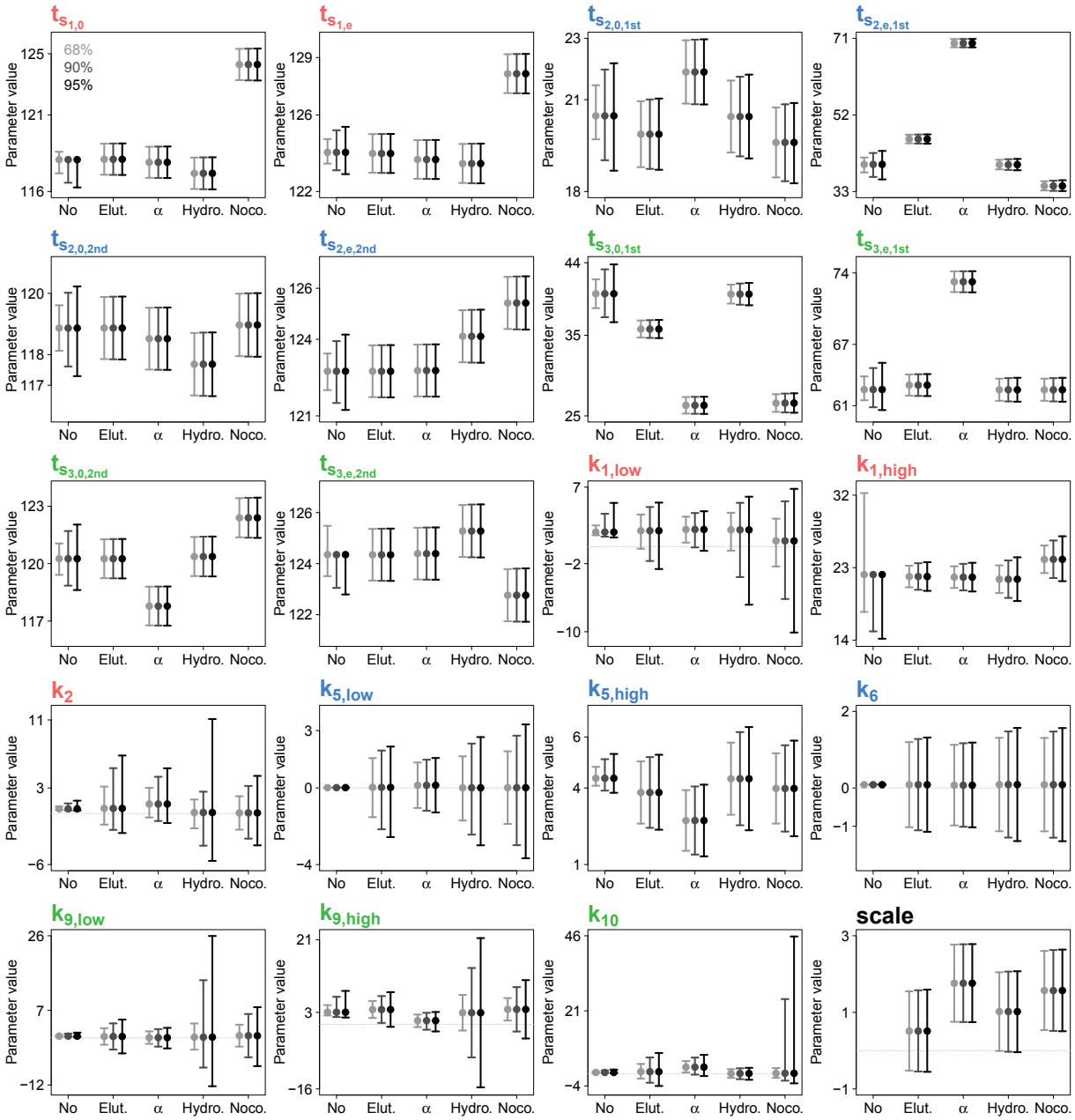

Figure E.9: Estimated mRNA parameters with confidence intervals of different confidence levels. In this figure, we show estimated mRNA parameters as well as the scaling factor introduced to combine mRNA and protein optimization (see Tables 4.2 and 4.1 for associated reactions). mRNA parameters are estimated from smFISH data in the mRNA optimization step and marked as "No" (abbreviation for "No synchronization"). mRNA parameters are re-estimated from Western blot data in the protein optimization step and assigned to respective synchronization methods: "Elut." (abbreviation for "Elutriation"), "α" (abbreviation for "α-factor"), "Hydro." (abbreviation for "Hydroxyurea") and "Noco." (abbreviation for "Nocodazole"). The scaling factor is estimated in the protein optimization step only and , therefore, has no estimate for "No". Error bars represent 95%, 90% and 68% confidence intervals calculated from profile likelihoods. If confidence intervals were not determinable for both sides, we plotted an one-sided open confidence interval. A gray dotted line indicates zero.

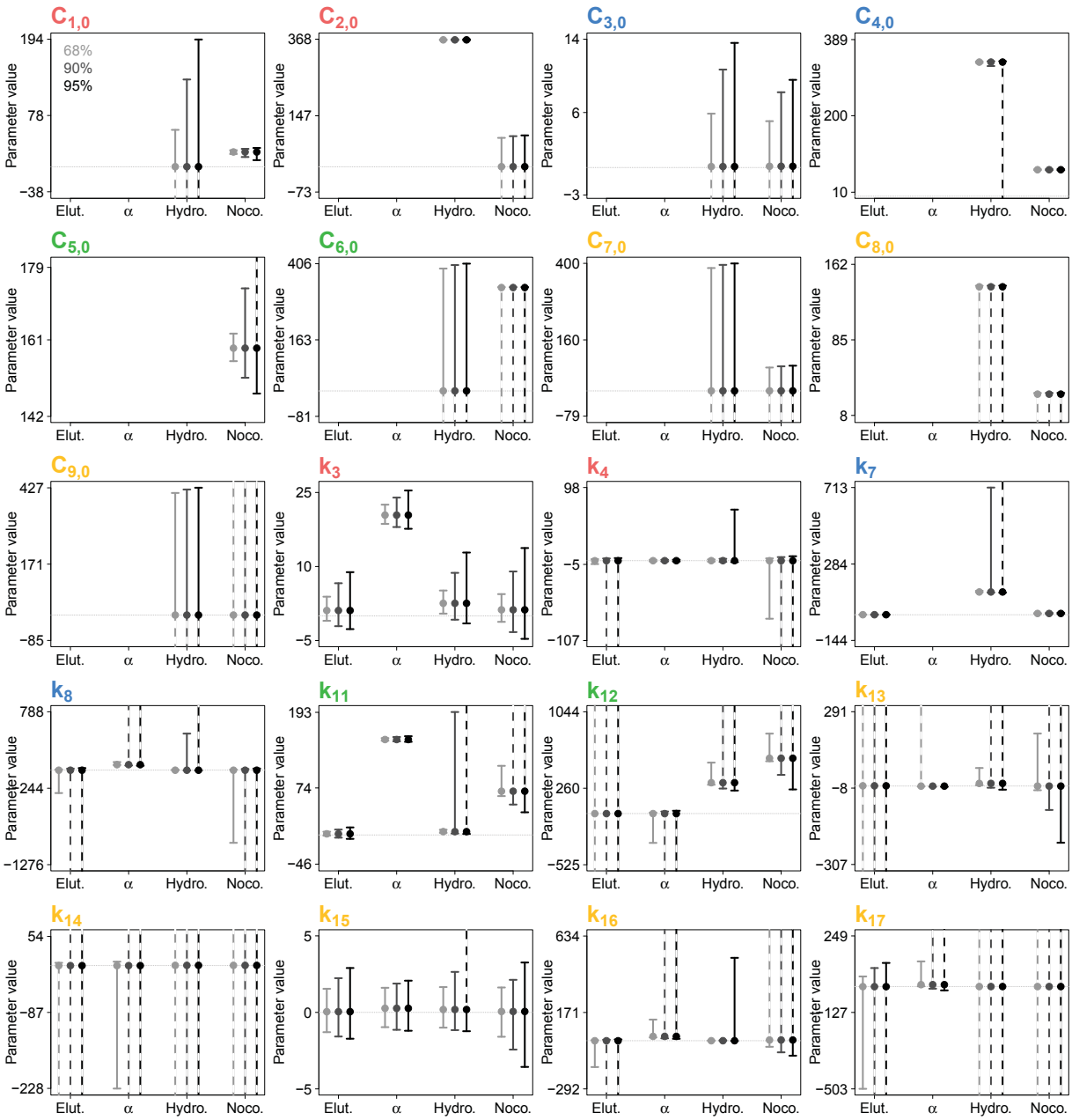

Figure E.10: Estimated protein parameters before L_1 regularization with confidence intervals of different confidence levels. In this figure, we show protein parameters (see Tables 4.2 and 4.1 for associated reactions) estimated from Western blot data in the protein optimization step. Parameter estimates are assigned to the respective synchronization method: "Elut." (abbreviation for "Elutriation"), "α" (abbreviation for "α-factor"), "Hydro." (abbreviation for "Hydroxyurea") and "Noco." (abbreviation for "Nocodazole"). Initial values are estimated for synchronization by hydroxyurea and nocodazole only. Parameters k_7 and $C_{5,0}$ are not missing for α-factor and hydroxyurea synchronization but their values are larger than the others. Error bars represent 95%, 90% and 68% confidence intervals calculated from profile likelihoods. Dashed lines instead of error bars indicate infinite confidence intervals. The gray dotted line indicates zero.

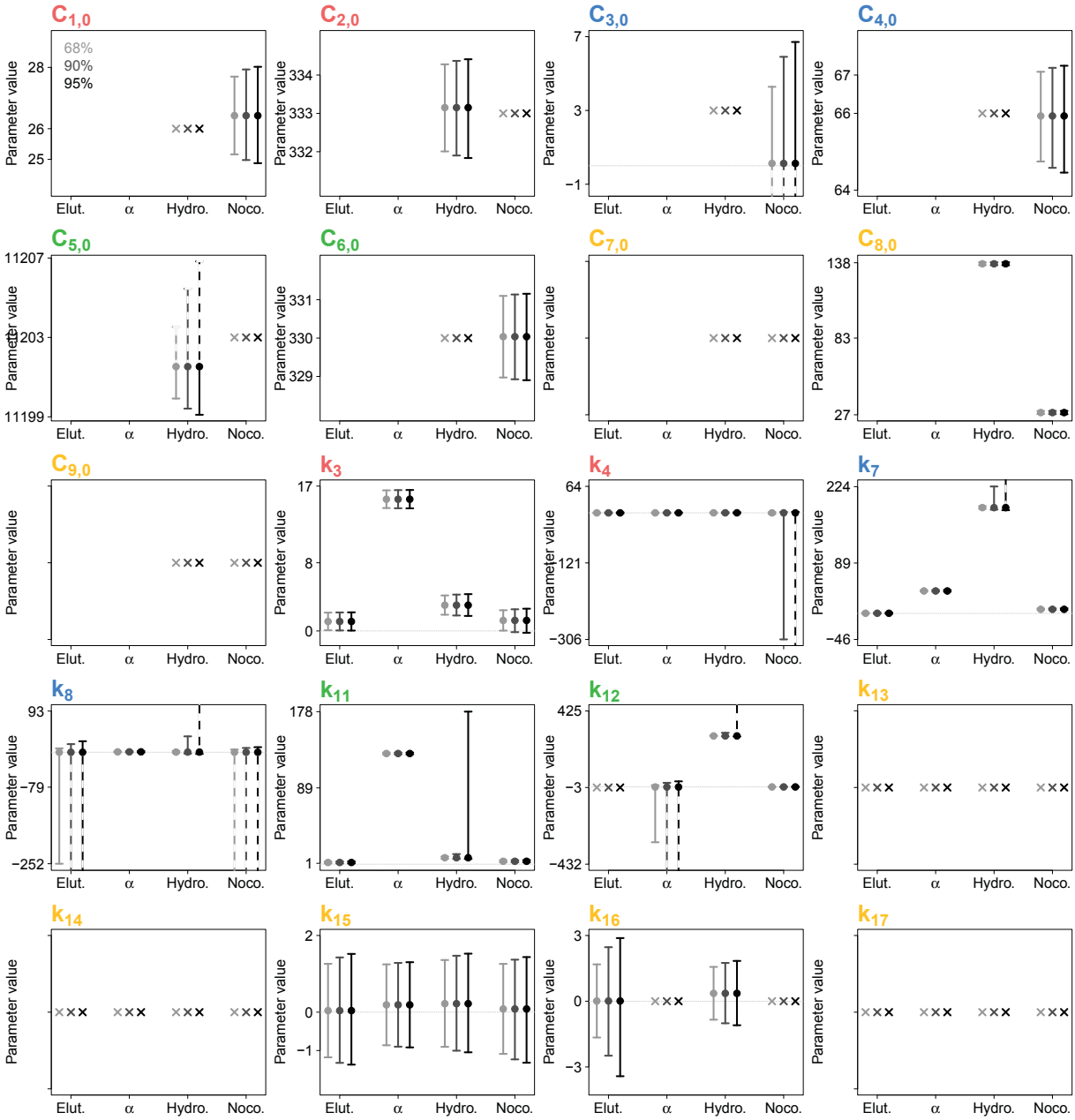

Figure E.11: Estimated protein parameters after L_1 regularization with confidence intervals of different confidence levels. In this figure, we show protein parameters (see Tables 4.2 and 4.1 for associated reactions) estimated from Western blot data in the protein optimization step after applying L_1 regularization. Removed parameters show a cross instead of a dot with error bars. Parameter estimates are assigned to the respective synchronization method: "Elut." (abbreviation for "Elutriation"), "α" (abbreviation for "α-factor"), "Hydro." (abbreviation for "Hydroxyurea") and "Noco." (abbreviation for "Nocodazole"). Initial values are estimated for synchronization by hydroxyurea and nocodazole only. Parameters k_7 and $C_{5,0}$ are not missing for α-factor and hydroxyurea synchronization but their values are larger than the others. Error bars represent 95%, 90% and 68% confidence intervals calculated from profile likelihoods. Dashed lines instead of error bars indicate infinite confidence intervals. The gray dotted line indicates zero.

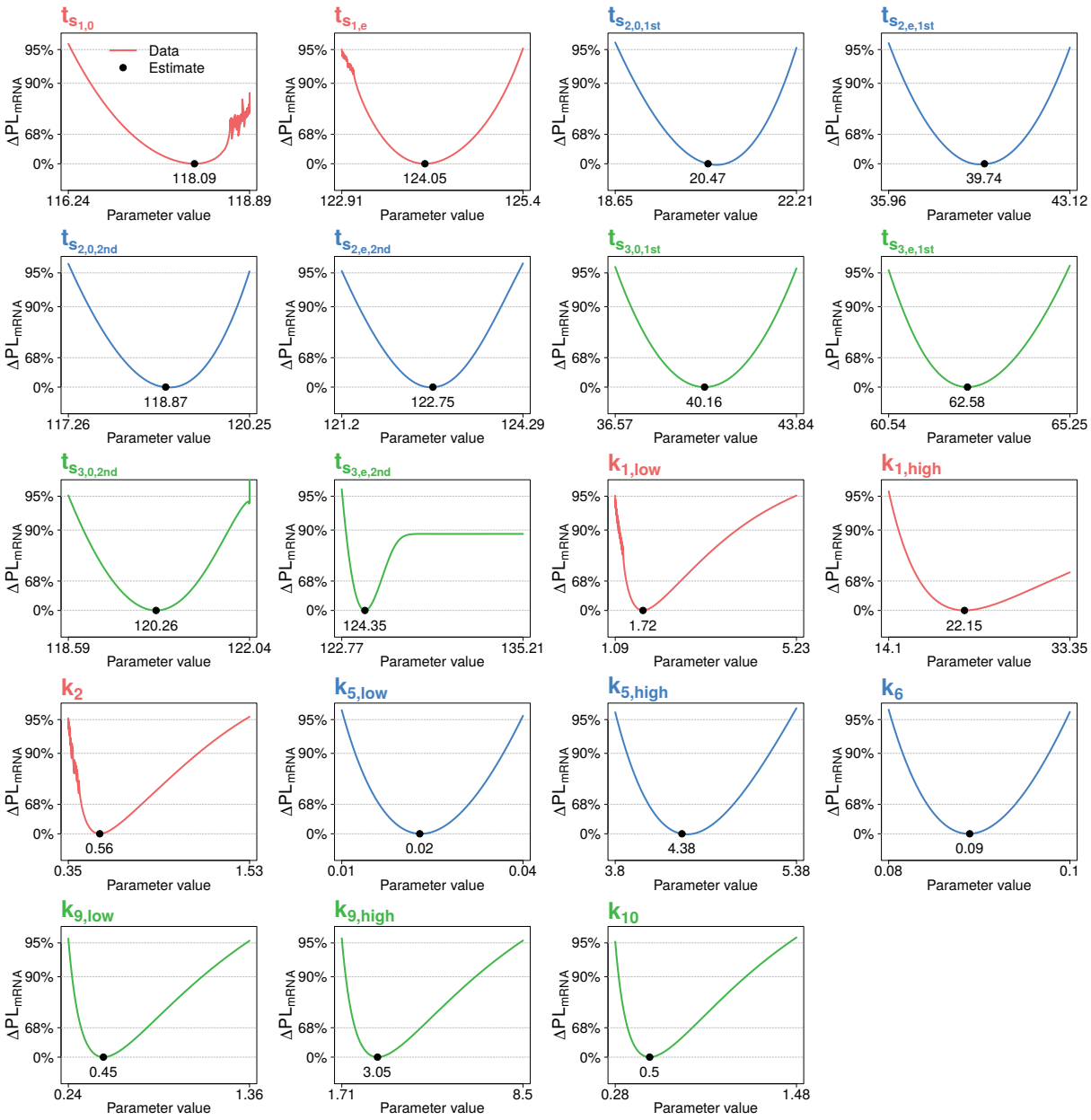

Figure E.12: Profile likelihoods of mRNA parameters estimated for unsynchronized cells in the mRNA optimization step. In this figure, we represent profile likelihoods of mRNA parameters (see Tables 4.2 and 4.1 for associated reactions) estimated from smFISH data in the mRNA optimization step as shown in Figure 6.1. Smallest, largest and estimated parameter values are plotted on the x-axis. Likelihood ratios and different confidence levels are given on the y-axis. Colored lines show contributions of the data. Red, blue and green refer to *SIC1*, *CLN2* and *CLB5*, respectively. Parameter estimates are marked as black dots.

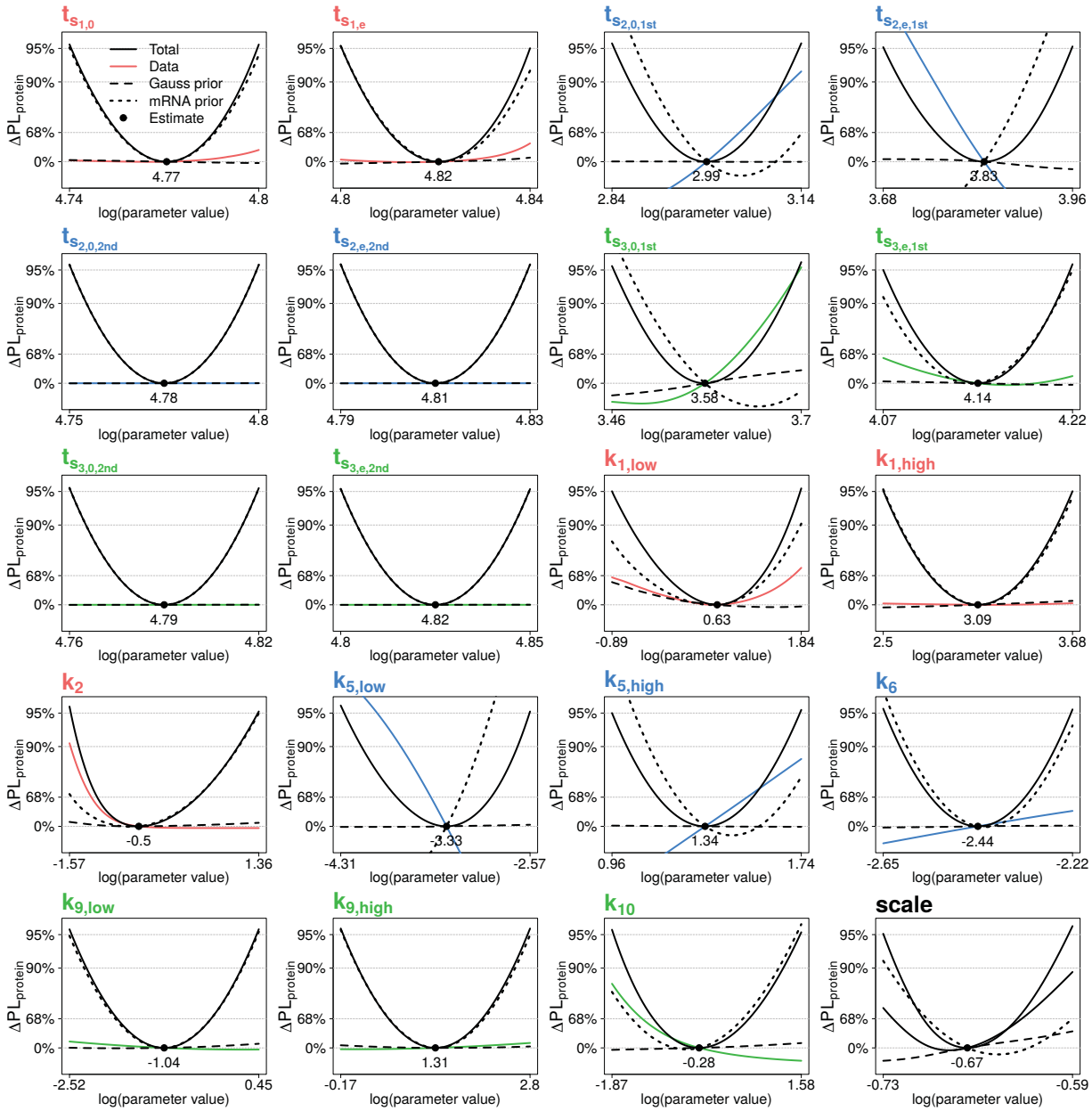

Figure E.13: Profile likelihoods of mRNA parameters re-estimated for elutriated cells in the protein optimization step. In this figure, we represent profile likelihoods of mRNA parameters and the scaling factor introduced to combine mRNA and protein optimization (see Tables 4.2 and 4.1 for associated reactions) estimated from Western blot data in the protein optimization step as shown in Figure 6.3. Smallest, largest and estimated parameter log-transformed values are plotted on the x-axis. Likelihood ratios and different confidence levels are given on the y-axis. Colored lines show contributions of the data. Red, blue and green refer to Sic1, Cln2 and Clb5, respectively. Contributions of the Gaussian and mRNA prior are illustrated with dashed and dotted black lines. The total profile is given by a black solid line. Parameter estimates are marked as black dots.

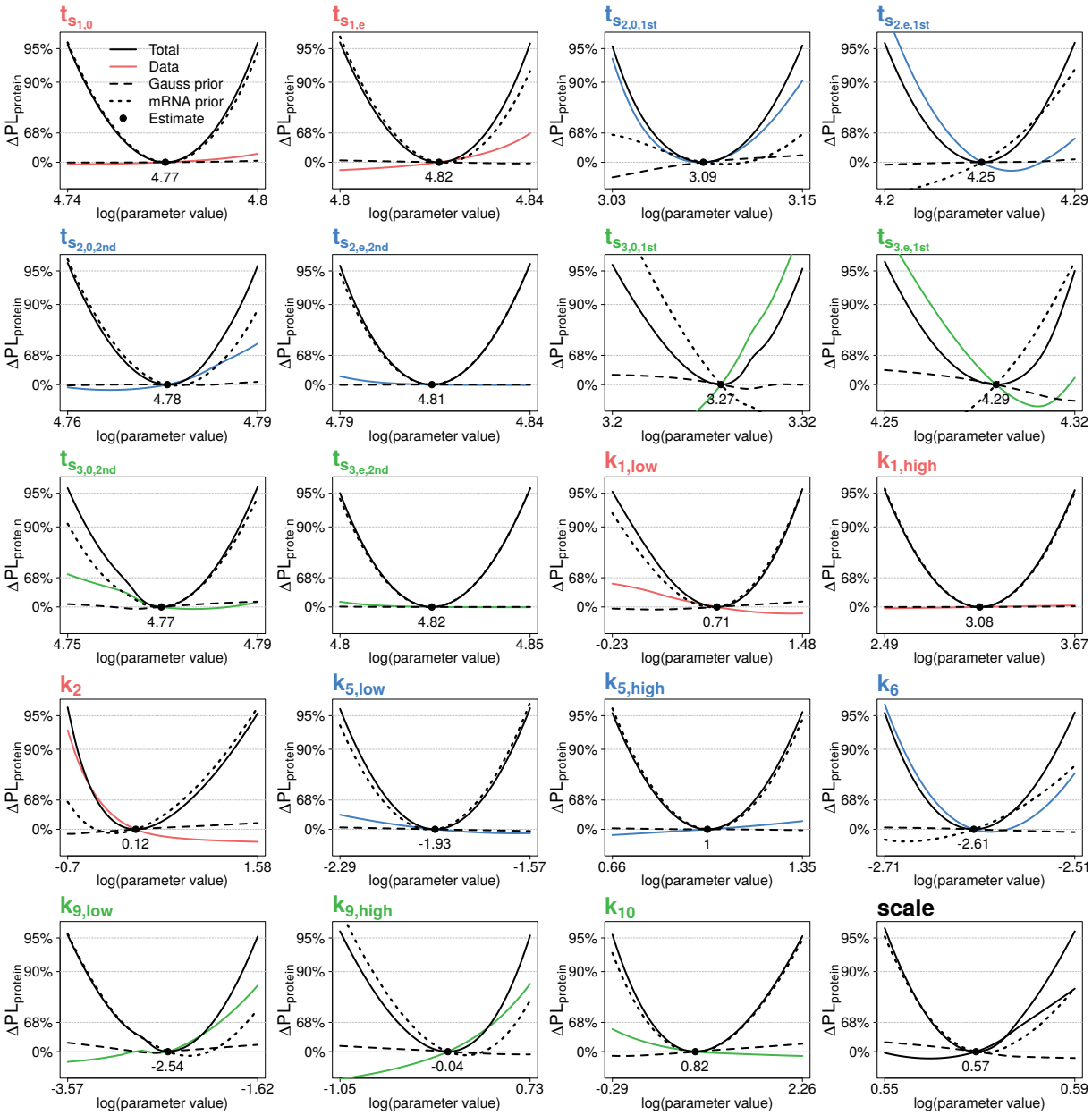

Figure E.14: Profile likelihoods of mRNA parameters re-estimated for α-factor synchronized cells in the protein optimization step. In this figure, we represent profile likelihoods of mRNA parameters and the scaling factor introduced to combine mRNA and protein optimization (see Tables 4.2 and 4.1 for associated reactions) estimated from Western blot data in the protein optimization step as shown in Figure 6.3. Smallest, largest and estimated log-transformed parameter values are plotted on the x-axis. Likelihood ratios and different confidence levels are given on the y-axis. Colored lines show contributions of the data. Red, blue and green refer to Sic1, Cln2 and Clb5, respectively. Contributions of the Gaussian and mRNA prior are illustrated with dashed and dotted black lines. The total profile is given by a black solid line. Parameter estimates are marked as black dots.

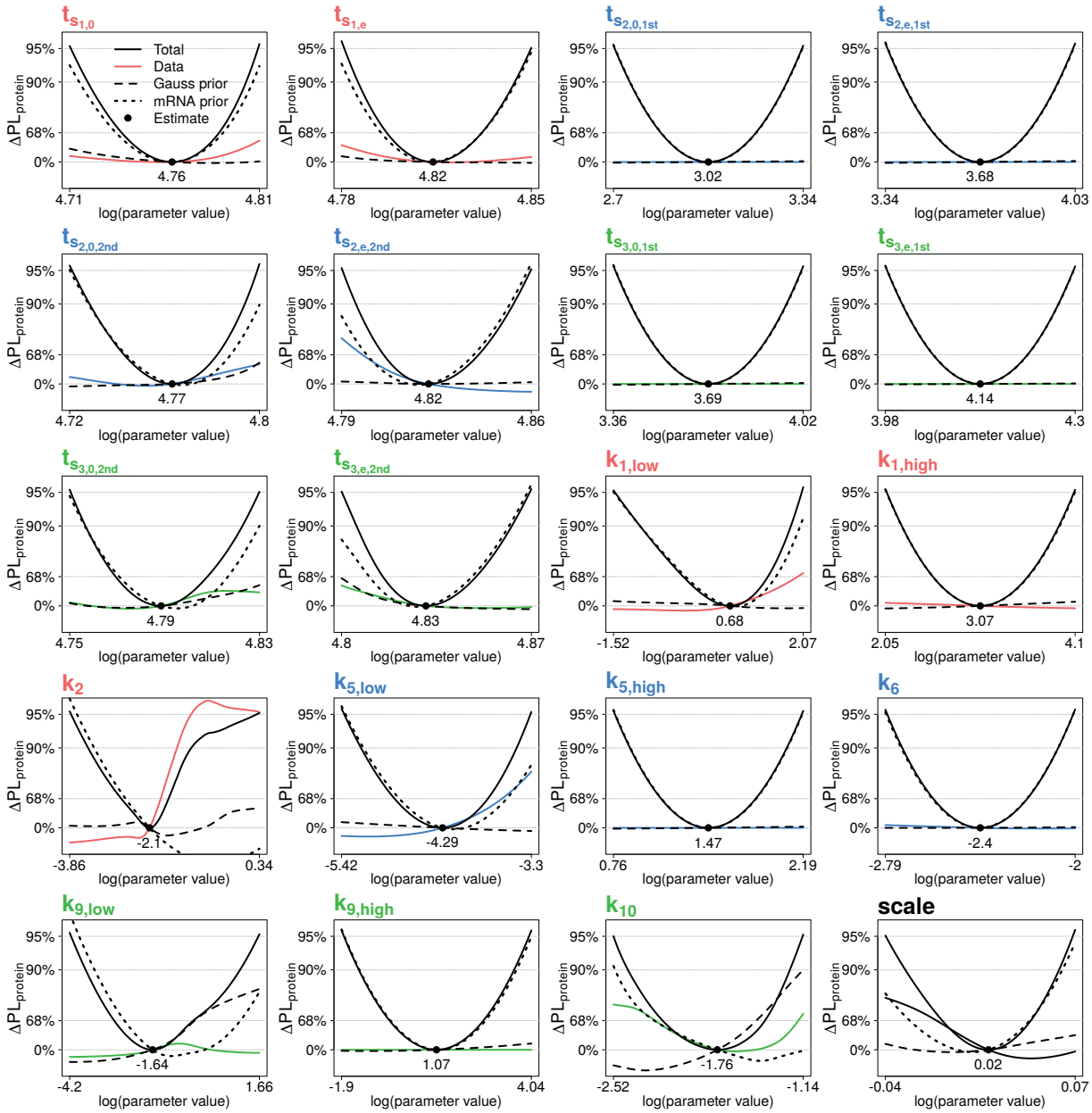

Figure E.15: Profile likelihoods of mRNA parameters re-estimated step for hydroxyurea synchronized cells in the protein optimization. In this figure, we represent profile likelihoods of mRNA parameters and the scaling factor introduced to combine mRNA and protein optimization (see Tables 4.2 and 4.1 for associated reactions) estimated from Western blot data in the protein optimization as shown in Figure 6.3. Smallest, largest and estimated log-transformed parameter values are plotted on the x-axis. Likelihood ratios and different confidence levels are given on the y-axis. Colored lines show contributions of the data. Red, blue and green refer to Sic1, Cln2 and Clb5, respectively. Contributions of the Gaussian and mRNA prior are illustrated with dashed and dotted black lines. The total profile is given by a black solid line. Parameter estimates are marked as black dots.

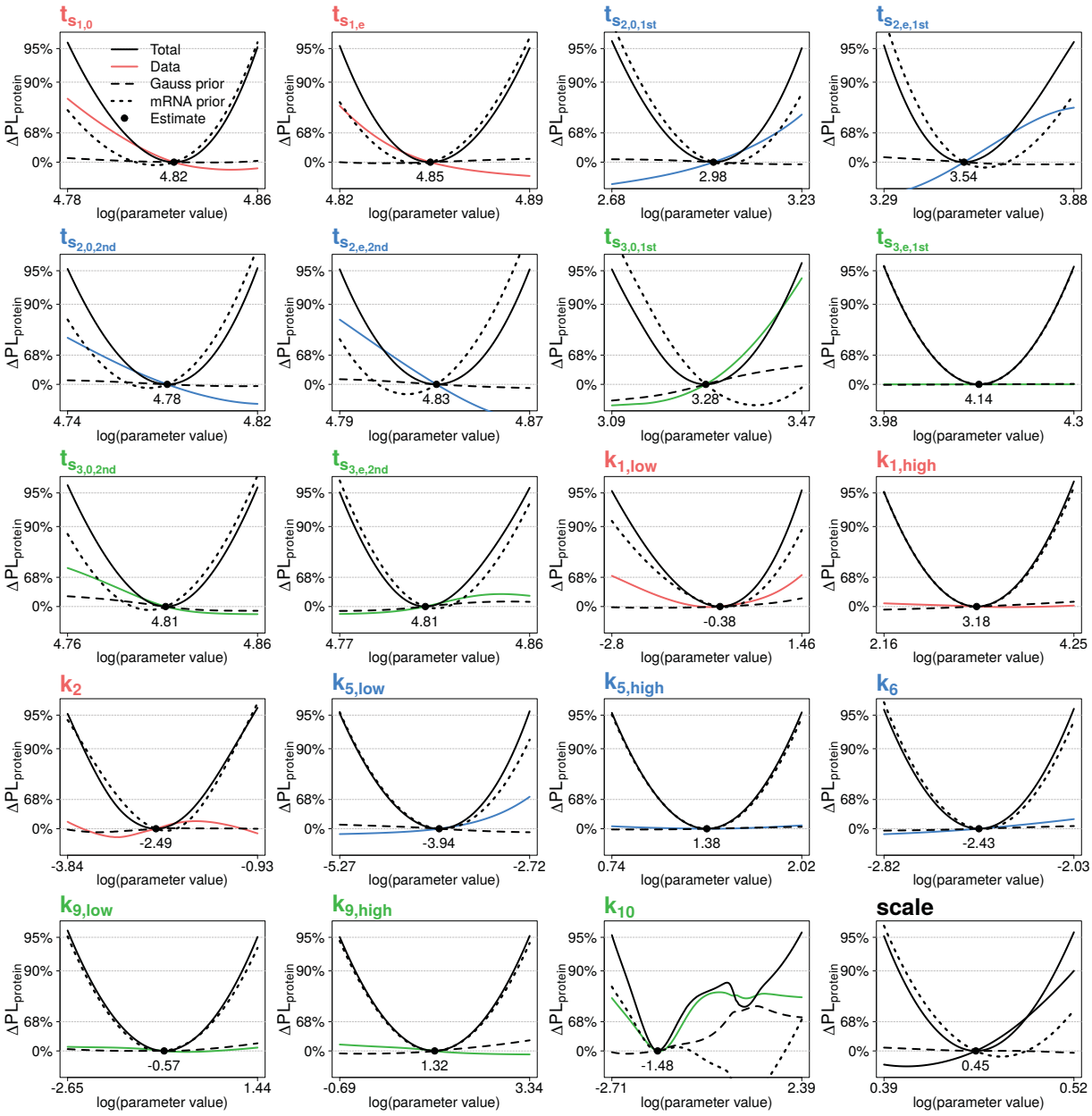

Figure E.16: Profile likelihoods of mRNA parameters re-estimated for nocodazole synchronized cells in the protein optimization. In this figure, we represent profile likelihoods of mRNA parameters and the scaling factor introduced to combine mRNA and protein optimization (see Tables 4.2 and 4.1 for associated reactions) estimated from Western blot data in the protein optimization as shown in Figure 6.3. Smallest, largest and estimated log-transformed parameter values are plotted on the x-axis. Likelihood ratios and different confidence levels are given on the y-axis. Colored lines show contributions of the data. Red, blue and green refer to Sic1, Cln2 and Clb5, respectively. Contributions of the Gaussian and mRNA prior are illustrated with dashed and dotted black lines. The total profile is given by a black solid line. Parameter estimates are marked as black dots.

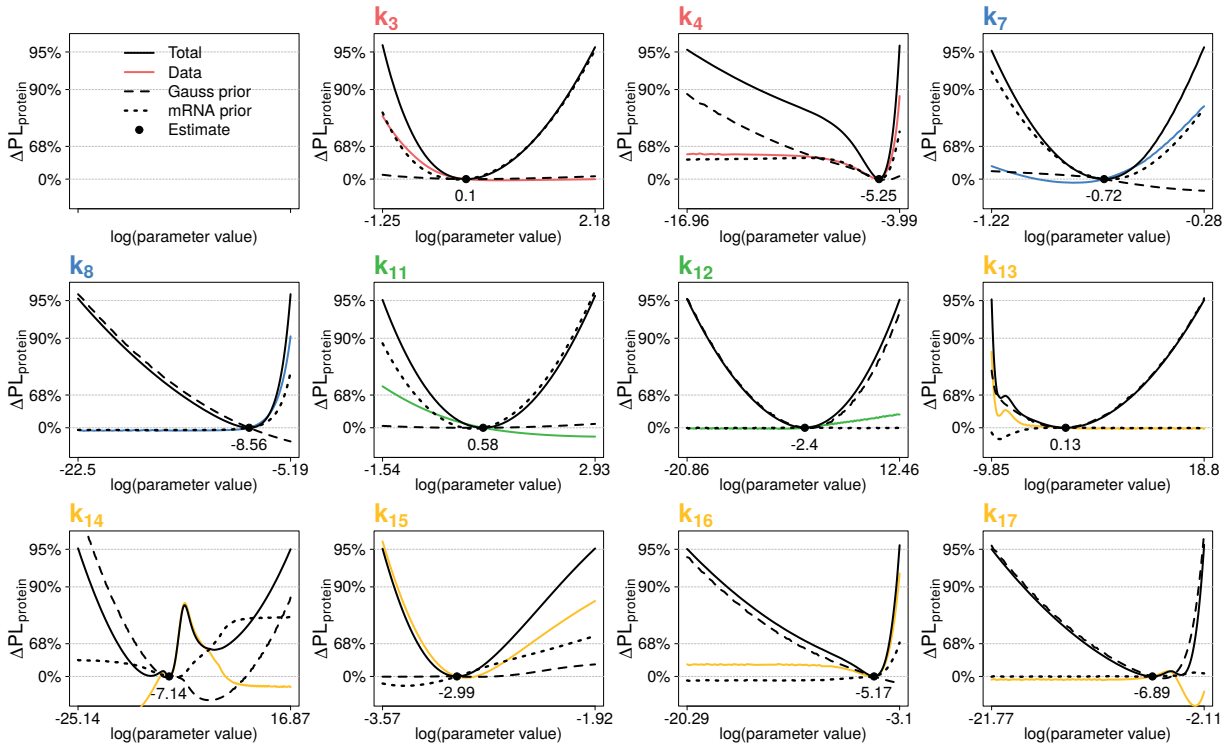

Figure E.17: Profile likelihoods of protein parameters estimated for elutriated cells in the protein optimization. In this figure, we represent profile likelihoods of protein parameters (see Tables 4.2 and 4.1 for associated reactions) estimated from Western blot data in the protein optimization as shown in Figure 6.3. Smallest, largest and estimated log-transformed parameter values are plotted on the x-axis. Likelihood ratios and different confidence levels are given on the y-axis. Colored lines show contributions of the data. Red, blue and green refer to Sic1, Cln2 and Clb5, respectively. Yellow represents a mixture of Sic1 and Clb5. Contributions of the Gaussian and mRNA prior are illustrated with dashed and dotted black lines. The total profile is given by a black solid line. Parameter estimates are marked as black dots.

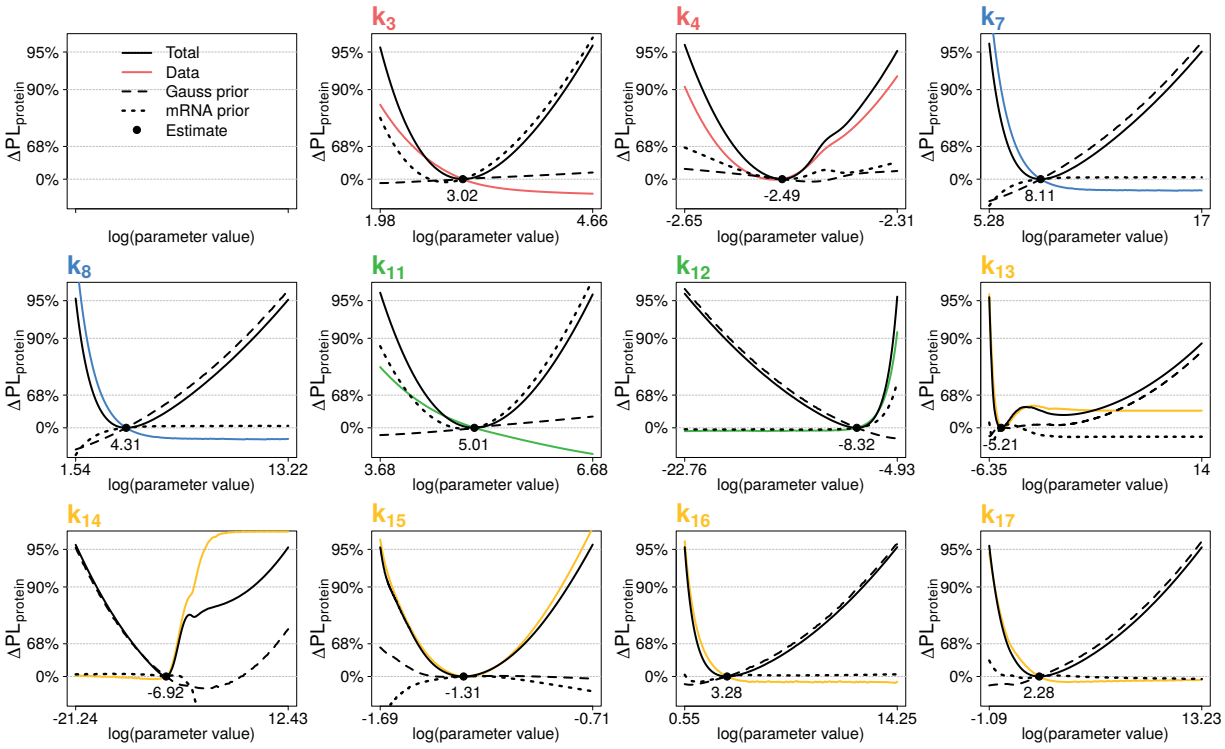

Figure E.18: Profile likelihoods of protein parameters estimated for α-factor synchronized cells in the protein optimization. In this figure, we represent profile likelihoods of protein parameters (see Tables 4.2 and 4.1 for associated reactions) estimated rom Western blot data in the protein optimization as shown in Figure 6.3. Smallest, largest and estimated log-transformed parameter values are plotted on the x-axis. Likelihood ratios and different confidence levels are given on the y-axis. Colored lines show contributions of the data. Red, blue and green refer to Sic1, Cln2 and Clb5, respectively. Yellow represents a mixture of Sic1 and Clb5. Contributions of the Gaussian and mRNA prior are illustrated with dashed and dotted black lines. The total profile is given by a black solid line. Parameter estimates are marked as black dots.

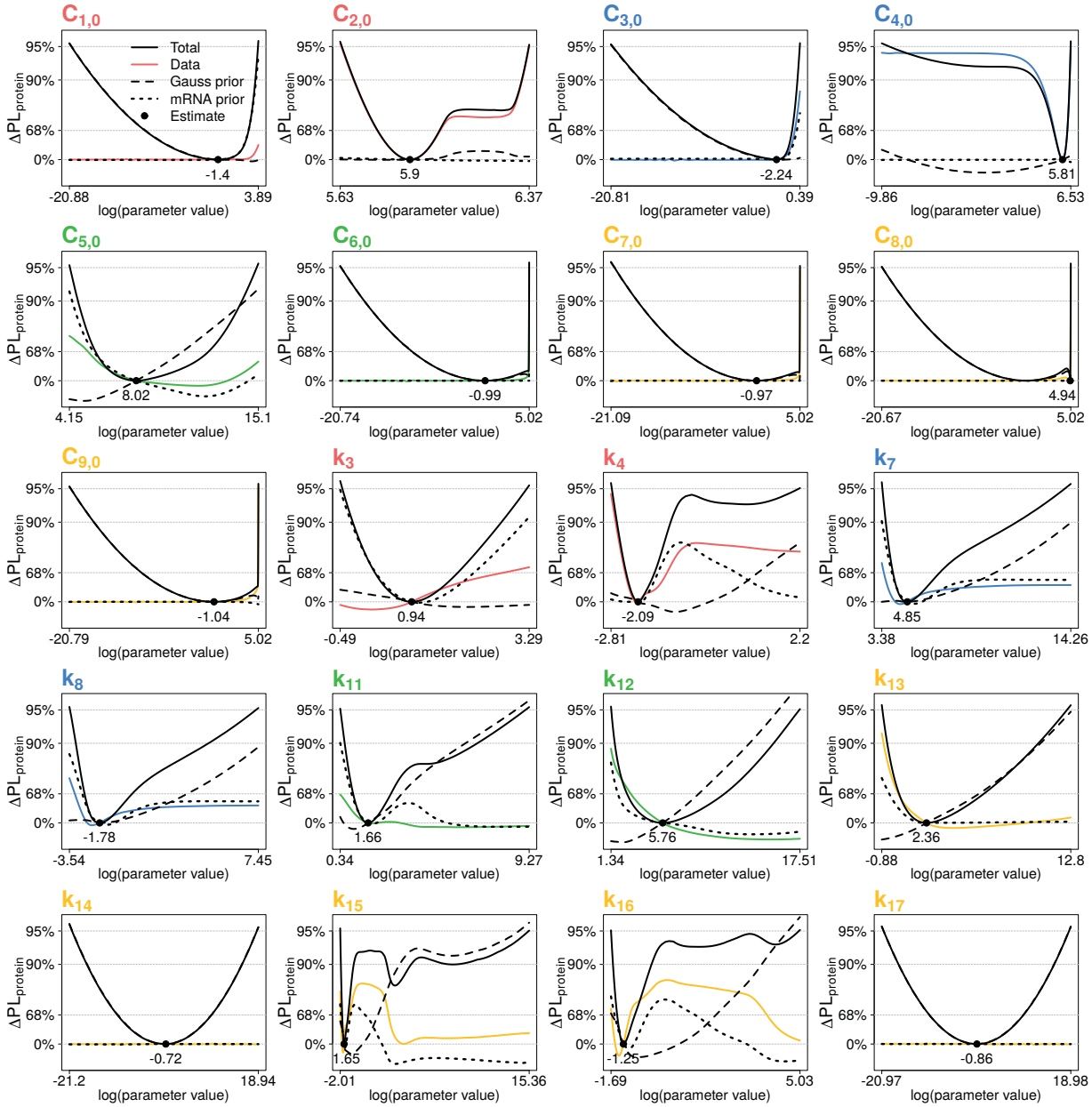

Figure E.19: Profile likelihoods of protein parameters estimated for hydroxyurea synchronized cells in the protein optimization. In this figure, we represent profile likelihoods of protein parameters (see Tables 4.2 and 4.1 for associated reactions) estimated from Western blot data in the protein optimization as shown in Figure 6.3. Smallest, largest and estimated log-transformed parameter values are plotted on the x-axis. Likelihood ratios and different confidence levels are given on the y-axis. Colored lines show contributions of the data. Red, blue and green refer to Sic1, Cln2 and Clb5, respectively. Yellow represents a mixture of Sic1 and Clb5. Contributions of the Gaussian and mRNA prior are illustrated with dashed and dotted black lines. The total profile is given by a black solid line. Parameter estimates are marked as black dots.

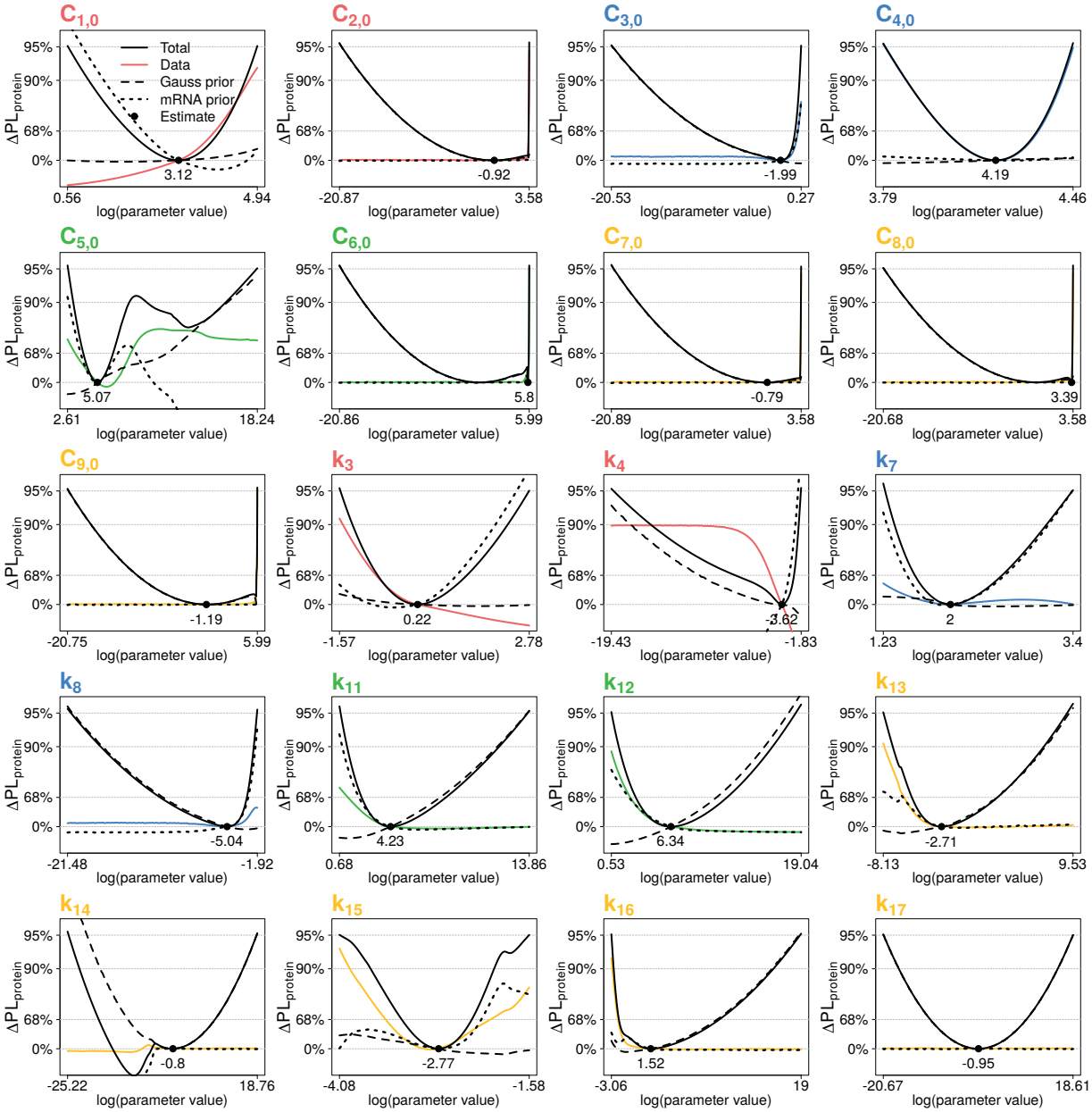

Figure E.20: Profile likelihoods of protein parameters estimated for nocodazole synchronized cells in the protein optimization. In this figure, we represent profile likelihoods of protein parameters (see Tables 4.2 and 4.1 for associated reactions) estimated from Western blot data in the protein optimization as shown in Figure 6.3. Smallest, largest and estimated log-transformed parameter values are plotted on the x-axis. Likelihood ratios and different confidence levels are given on the y-axis. Colored lines show contributions of the data. Red, blue and green refer to Sic1, Cln2 and Clb5, respectively. Yellow represents a mixture of Sic1 and Clb5. Contributions of the Gaussian and mRNA prior are illustrated with dashed and dotted black lines. The total profile is given by a black solid line. Parameter estimates are marked as black dots.

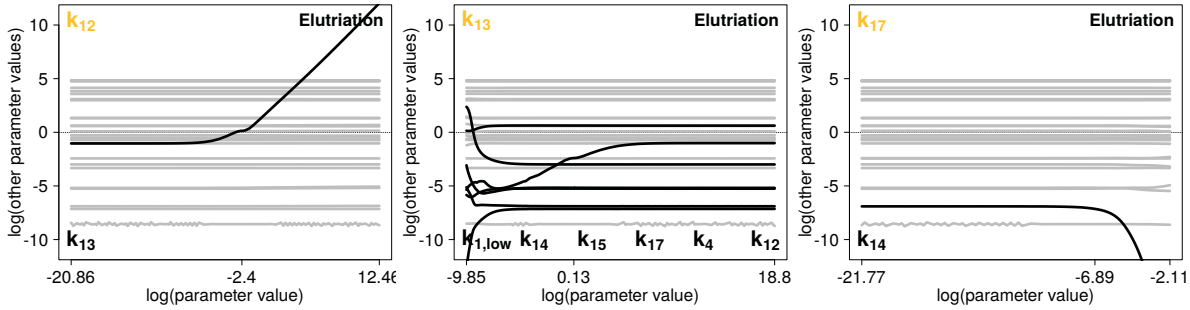

Figure E.21: Parameter values along profile likelihoods of structurally non-identifiable protein parameters for synchronization by elutriation. In this figure, we show parameter values along profile likelihoods of protein parameters k_{12}, k_{13} and k_{17} which are determined as being structurally non-identifiable. Parameter values are log-transformed. Gray and black lines indicate independent and dependent parameters, respectively. Dependent parameters are noted in the plot. A detailed plot description is given in Figure 6.2.

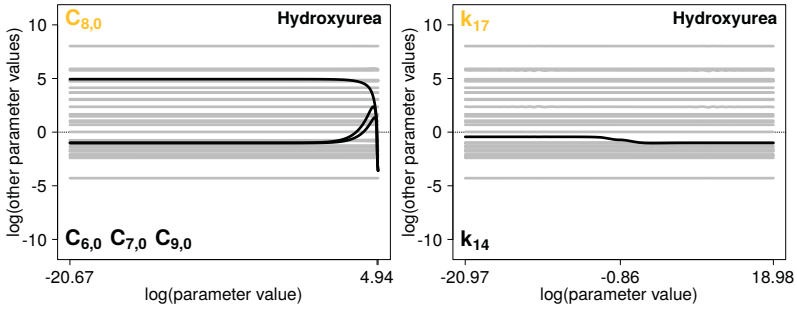

Figure E.22: Parameter values along profile likelihoods of structurally non-identifiable protein parameters for synchronization by hydroxyurea. In this figure, we show parameter values along profile likelihoods of protein parameters $C_{8,0}$ and k_{17} which are determined as being structurally non-identifiable. Parameter values are log-transformed. Gray and black lines indicate independent and dependent parameters, respectively. Dependent parameters are noted in the plot. A detailed plot description is given in Figure 6.2.

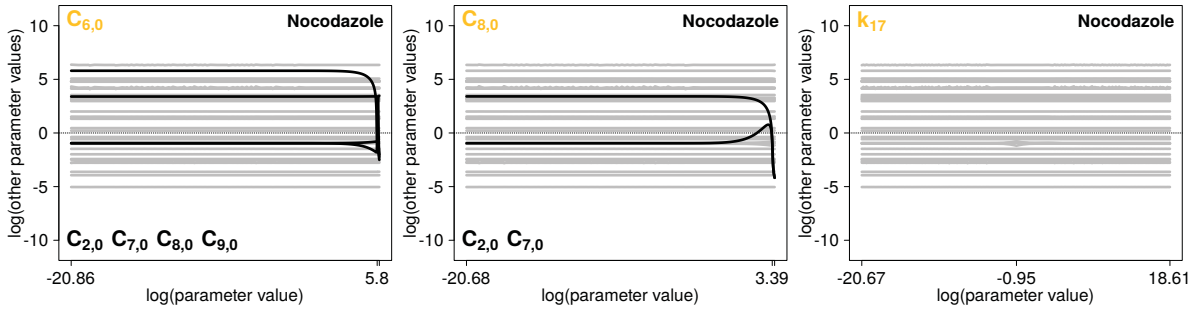

Figure E.23: Parameter values along profile likelihoods of structurally non-identifiable protein parameters for synchronization by nocodazole. In this figure, we show parameter values along profile likelihoods of protein parameters $C_{6,0}$, $C_{8,0}$ and k_{17} which are determined as being structurally non-identifiable. Parameter values are log-transformed. Gray and black lines indicate independent and dependent parameters, respectively. Dependent parameters are noted in the plot. A detailed plot description is given in Figure 6.2.

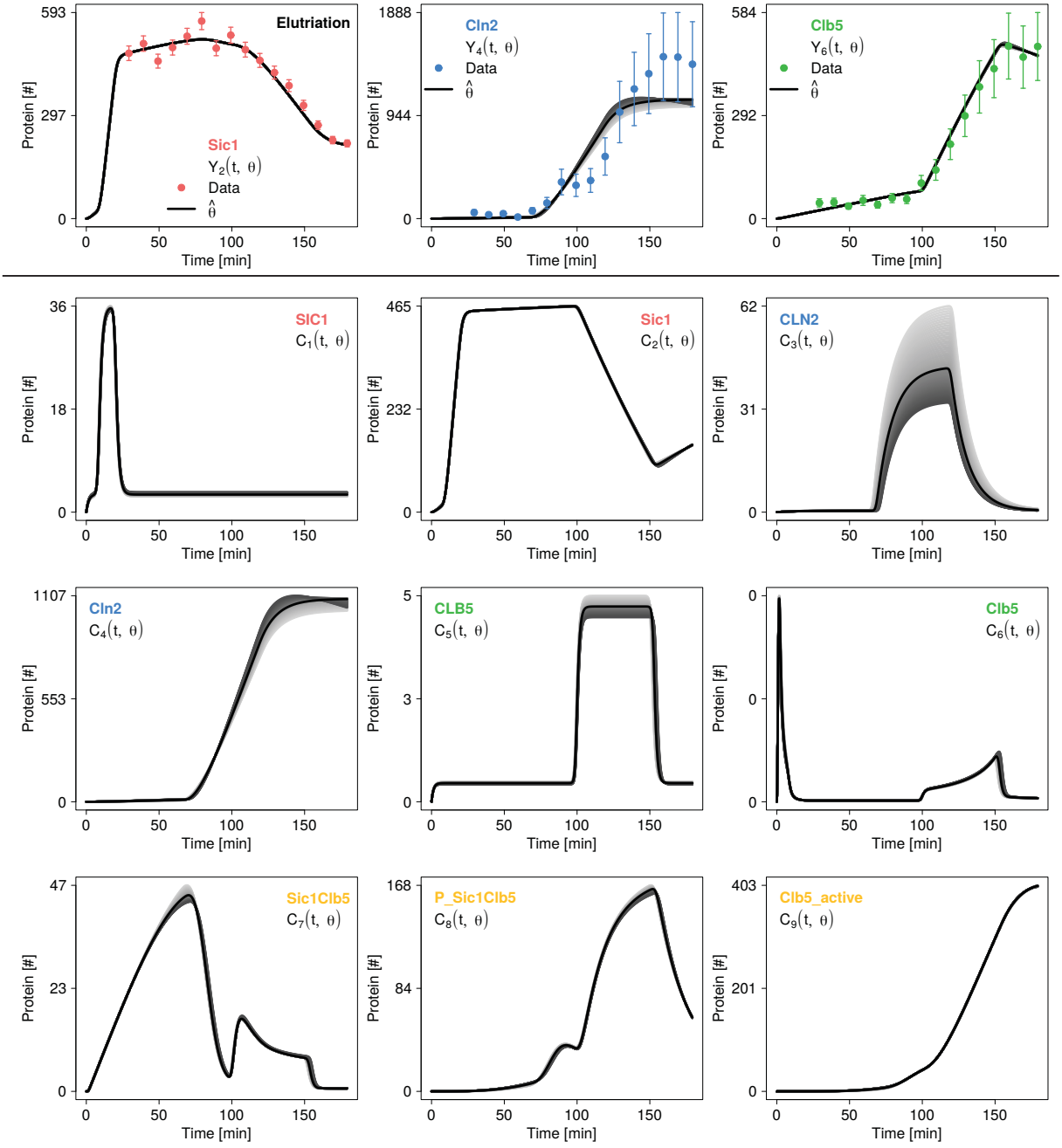

Figure E.24: Trajectories along the profile likelihood of protein parameter k_7 estimated for elutriated cells in the protein optimization step . In this figure, we plot trajectories of observables (above black horizontal line) and variables (below black horizontal line) along the profile likelihood of protein parameter k_7 estimated for elutriated cells in the protein optimization step. Observables are plotted together with the data (colored dots and error bars). Trajectories go from smallest (light gray) to largest (dark gray) parameter values. Trajectories for parameter optima are given as black lines. Red, blue and green species names refer to *SIC1*/Sic1, *CLN2*/Cln2 and *CLB5*/Clb5, respectively as in the data. Yellow represents a mixture of Sic1 and Clb5.

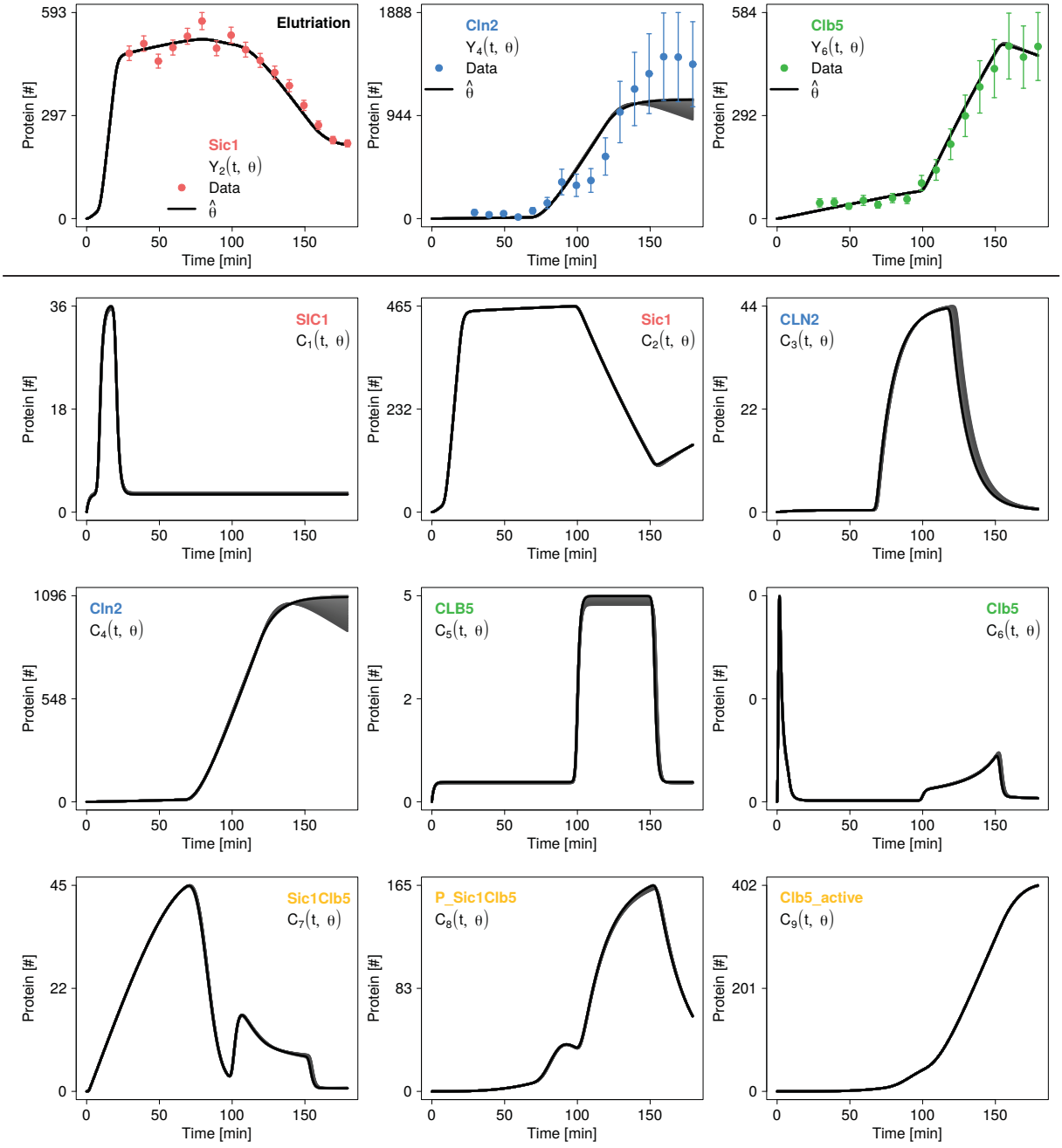

Figure E.25: Trajectories along the profile likelihood of protein parameter k_8 estimated for elutriated cells in the protein optimization step. In this figure, we plot trajectories of observables (above black horizontal line) and variables (below black horizontal line) along the profile likelihood of protein parameter k_8 estimated for elutriated cells in the protein optimization step. Observables are plotted together with the data (colored dots and error bars). Trajectories go from smallest (light gray) to largest (dark gray) parameter values. Trajectories for parameter optima are given as black lines. Red, blue and green species names refer to *SIC1*/Sic1, *CLN2*/Cln2 and *CLB5*/Clb5, respectively as in the data. Yellow represents a mixture of Sic1 and Clb5.

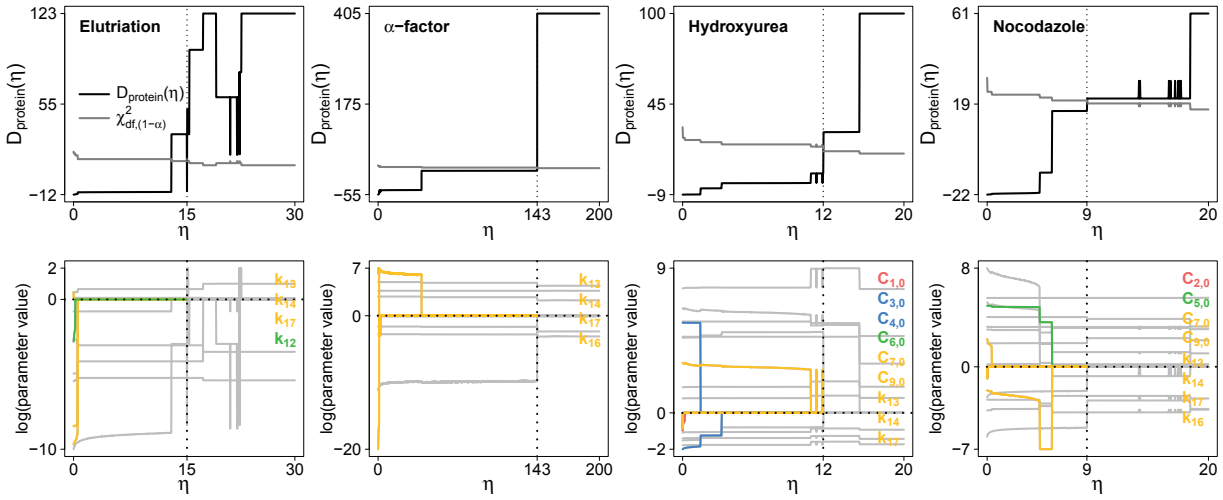

Figure E.26: Identification of the the optimal regularization strength in the L_1 regularization for protein parameters. In this figure, we show how the optimal regularization strength $\hat{\eta}$ is reached (upper row) and which parameters are finally non-selected (lower row) in the L_1 regularization for protein parameters of different synchronizations ("Elutriation", "α-factor", "Hydroxyurea" and "Nocodazole"). Non-selected parameter values are colored as long as the optimal regularization strength is reached. The optimal regularization strength (dotted line) is the largest value of η where the likelihood ratio $D_{protein}(\eta)$ (see Equation 7.8) is still below the $(1-\alpha)$ quantile of the $\chi^2(df)$ distribution with significance level $\alpha = 0.05$.

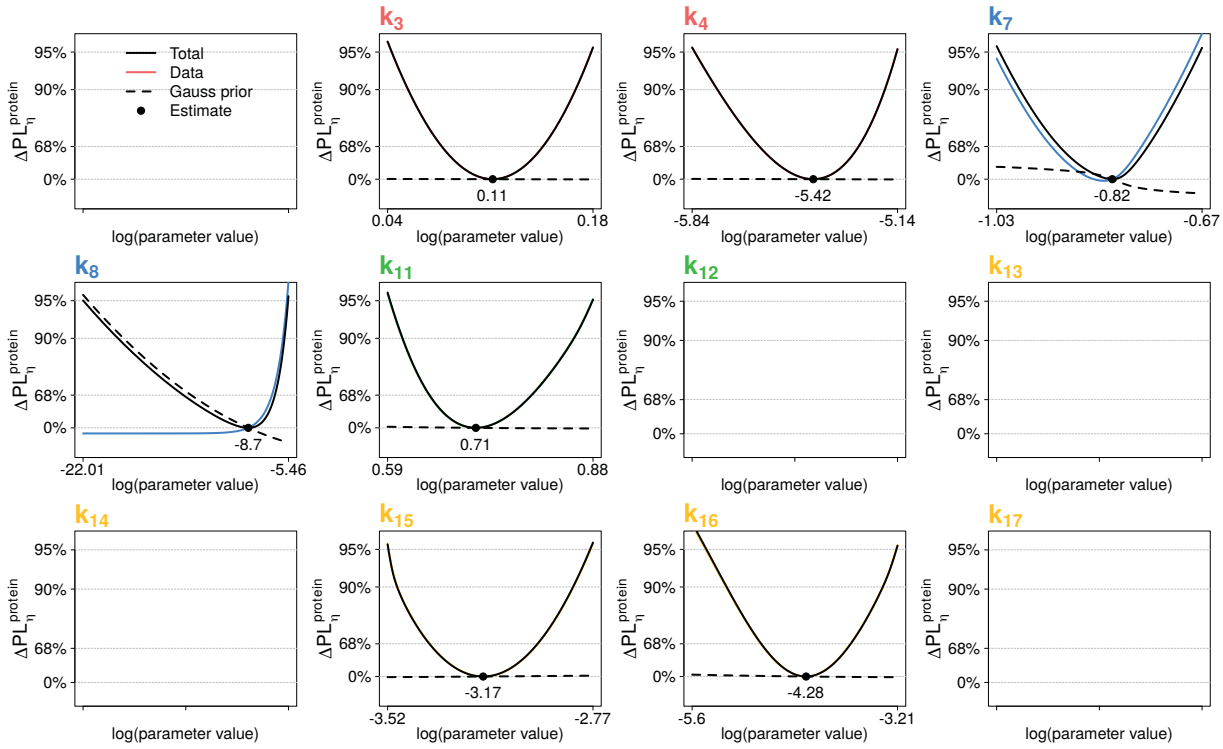

Figure E.27: Profile likelihoods of protein parameters estimated for elutriated cells after L_1 regularization in the protein optimization step. In this figure, we represent profile likelihoods of protein parameters (see Tables 4.2 and 4.1 for associated reactions) estimated from Western blot data after applying L_1 regularization to protein parameters in the protein optimization step as shown in Figure 7.1. Smallest, largest and estimated log-transformed parameter values are plotted on the x-axis. Likelihood ratios and different confidence levels are given on the y-axis. Colored lines show contributions of the data. Red, blue and green refer to Sic1, Cln2 and Clb5, respectively. Yellow represents a mixture of Sic1 and Clb5. The contribution of the Gaussian prior is given as dashed line and the total profile as black solid line. Parameter estimates are marked as black dots. An empty plot shows parameters which gone to zero during L_1 regularization.

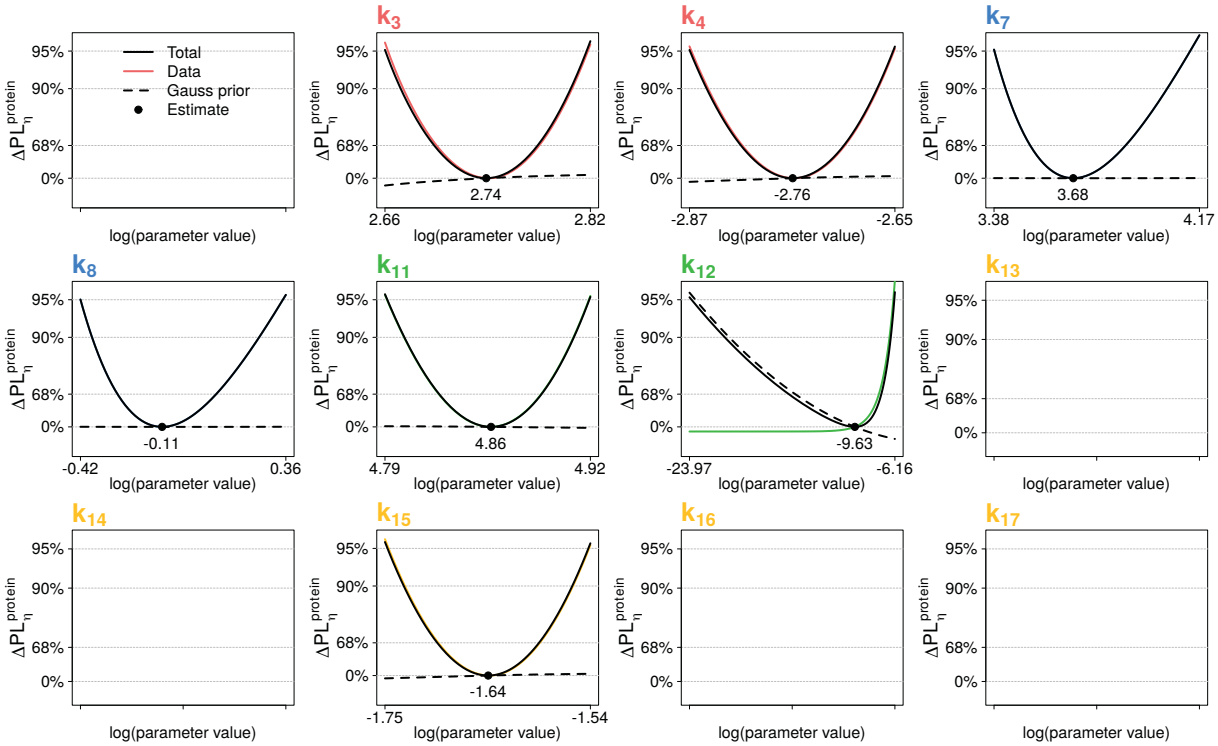

Figure E.28: Profile likelihoods of protein parameters estimated for α-factor synchronized cells after L_1 regularization in the protein optimization step. In this figure, we represent profile likelihoods of protein parameters (see Tables 4.2 and 4.1 for associated reactions) estimated from Western blot data after applying L_1 regularization to protein parameters in the protein optimization step as shown in Figure 7.1. Smallest, largest and estimated log-transformed parameter values are plotted on the x-axis. Likelihood ratios and different confidence levels are given on the y-axis. Colored lines show contributions of the data. Red, blue and green refer to Sic1, Cln2 and Clb5, respectively. Yellow represents a mixture of Sic1 and Clb5. The contribution of the Gaussian prior is given as dashed line and the total profile as black solid line. Parameter estimates are marked as black dots. An empty plot shows parameters which gone to zero during L_1 regularization.

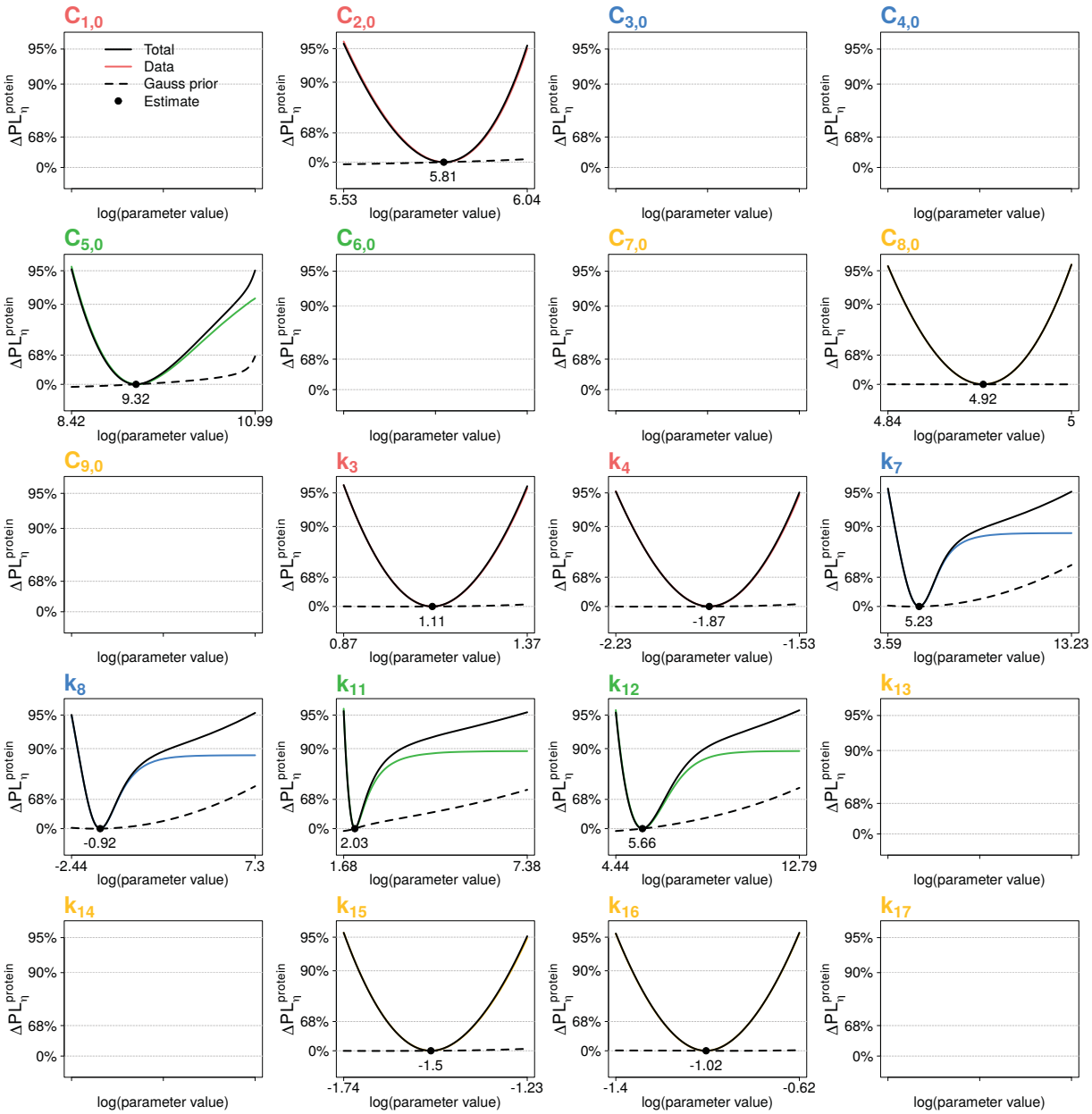

Figure E.29: Profile likelihoods of protein parameters estimated for hydroxyurea synchronized cells after L_1 regularization in the protein optimization step. In this figure, we represent profile likelihoods of protein parameters (see Tables 4.2 and 4.1 for associated reactions) estimated from Western blot data after applying L_1 regularization to protein parameters in the protein optimization step as shown in Figure 7.1. Smallest, largest and estimated log-transformed parameter values are plotted on the x-axis. Likelihood ratios and different confidence levels are given on the y-axis. Colored lines show contributions of the data. Red, blue and green refer to Sic1, Cln2 and Clb5, respectively. Yellow represents a mixture of Sic1 and Clb5. The contribution of the Gaussian prior is given as dashed line and the total profile as black solid line. Parameter estimates are marked as black dots. An empty plot shows parameters which gone to zero during L_1 regularization.

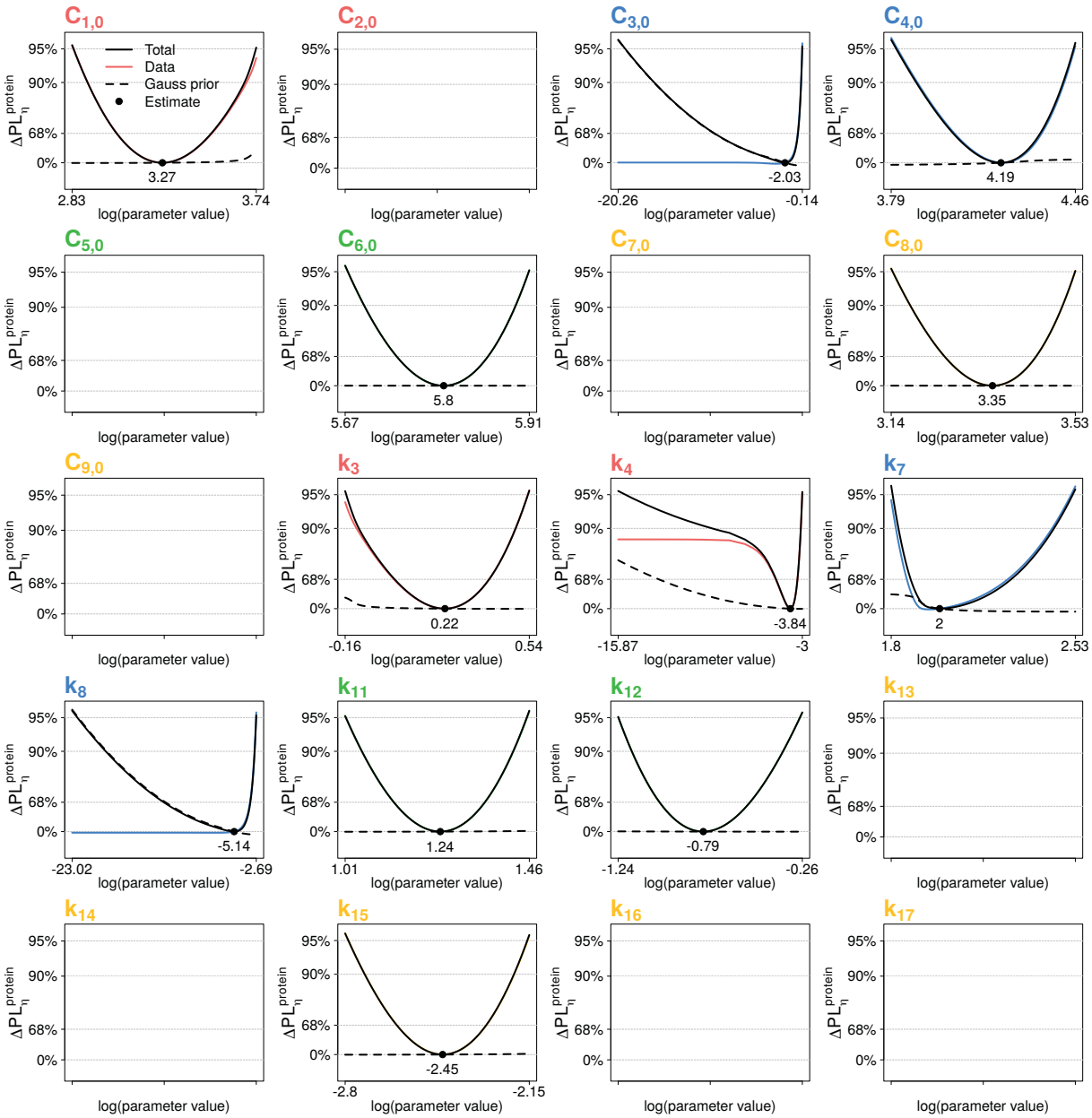

Figure E.30: Profile likelihoods of protein parameters estimated for nocodazole synchronized cells after L_1 regularization in the protein optimization step. In this figure, we represent profile likelihoods of protein parameters (see Tables 4.2 and 4.1 for associated reactions) estimated from Western blot data after applying L_1 regularization to protein parameters in the protein optimization step as shown in Figure 7.1. Smallest, largest and estimated log-transformed parameter values are plotted on the x-axis. Likelihood ratios and different confidence levels are given on the y-axis. Colored lines show contributions of the data. Red, blue and green refer to Sic1, Cln2 and Clb5, respectively. Yellow represents a mixture of Sic1 and Clb5. The contribution of the Gaussian prior is given as dashed line and the total profile as black solid line. Parameter estimates are marked as black dots. An empty plot shows parameters which gone to zero during L_1 regularization.

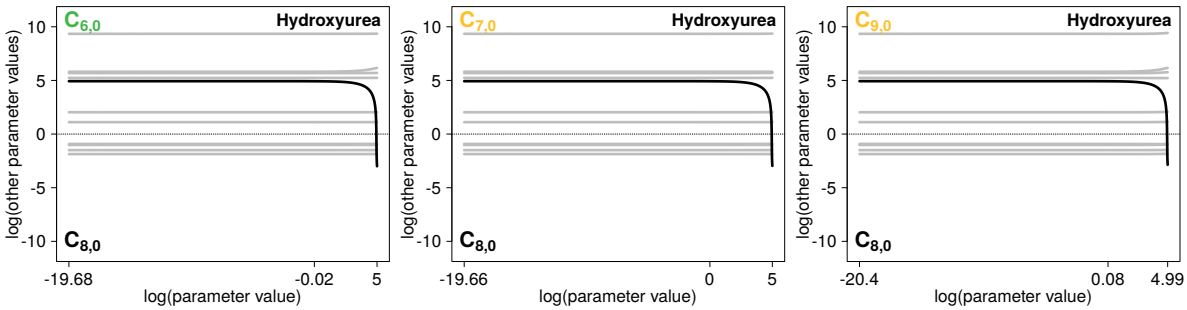

Figure E.31: Parameter values along the profile likelihoods of L_1 removed protein parameters for hydroxyurea synchronization. In this figure, we show parameter values along profile likelihoods of L_1 removed protein parameters $C_{6,0}$, $C_{7,0}$ and $C_{9,0}$ which are subsequently re-introduced to test uniqueness of the solution. Parameter values are log-transformed. Black lines indicate selected parameters which are exchangeable with the non-selected parameter why the non-selected parameter is not uniquely determined to be zero. Exchangeable parameters are noted in the plot. A detailed plot description is given in Figure 6.2.

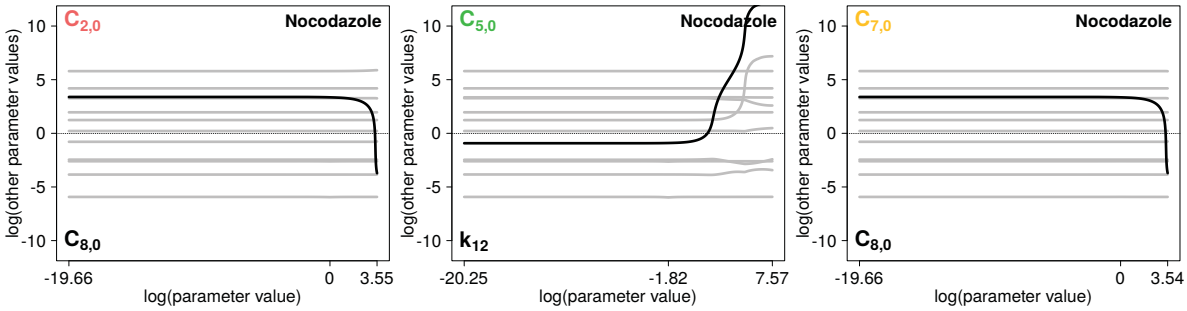

Figure E.32: Parameter values along profile likelihoods of L_1 removed protein parameters for nocodazole synchronization. In this figure, we show parameter values along profile likelihoods of L_1 removed protein parameters $C_{1,0}$, $C_{5,0}$ and $C_{7,0}$ which are subsequently re-introduced to test uniqueness of the solution. All parameter values are log-transformed. Black lines indicate selected parameters which are exchangeable with the non-selected parameter why the non-selected parameter is not uniquely determined to be zero. Exchangeable parameters are noted in the plot. A detailed plot description is given in Figure 6.2.

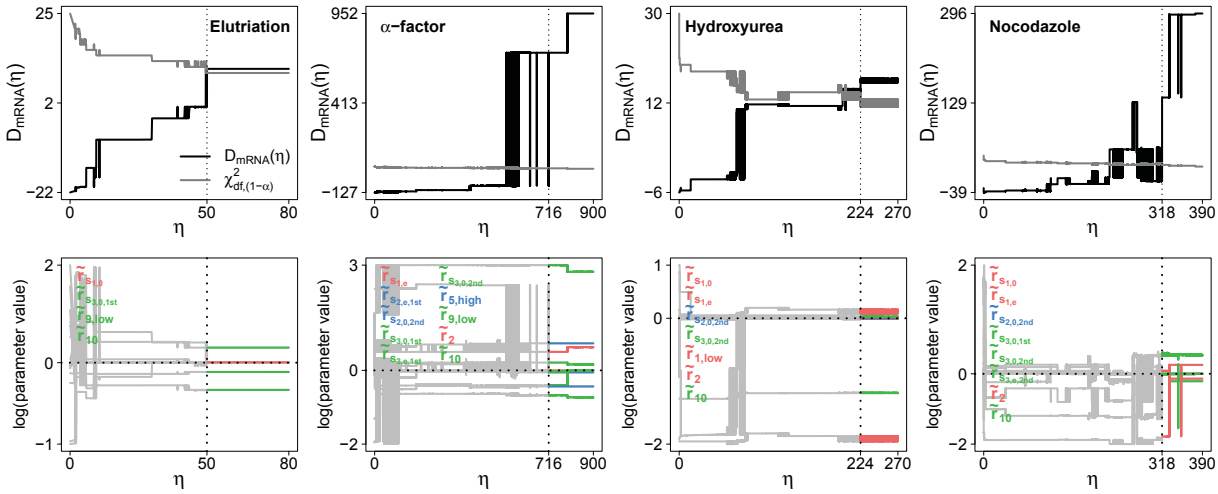

Figure E.33: Identification of the the optimal regularization strength in the L_1 regularization for mRNA fold changes. In this figure, we show how the optimal regularization strength $\hat{\eta}$ is reached (upper row) and which mRNA fold changes are finally selected (lower row) in the L_1 regularization for mRNA fold changes of different synchronizations ("Elutriation", "α-factor", "Hydroxyurea" and "Nocodazole"). Selected mRNA fold changes are colored after the optimal regularization strength is reached. The optimal regularization strength (vertical dotted line) is the largest value of η where the likelihood ratio $D_{mRNA}(\eta)$ (see Equation 7.11) is still below the $(1-\alpha)$ quantile of the $\chi^2(df)$ distribution with significance level $\alpha = 0.05$.

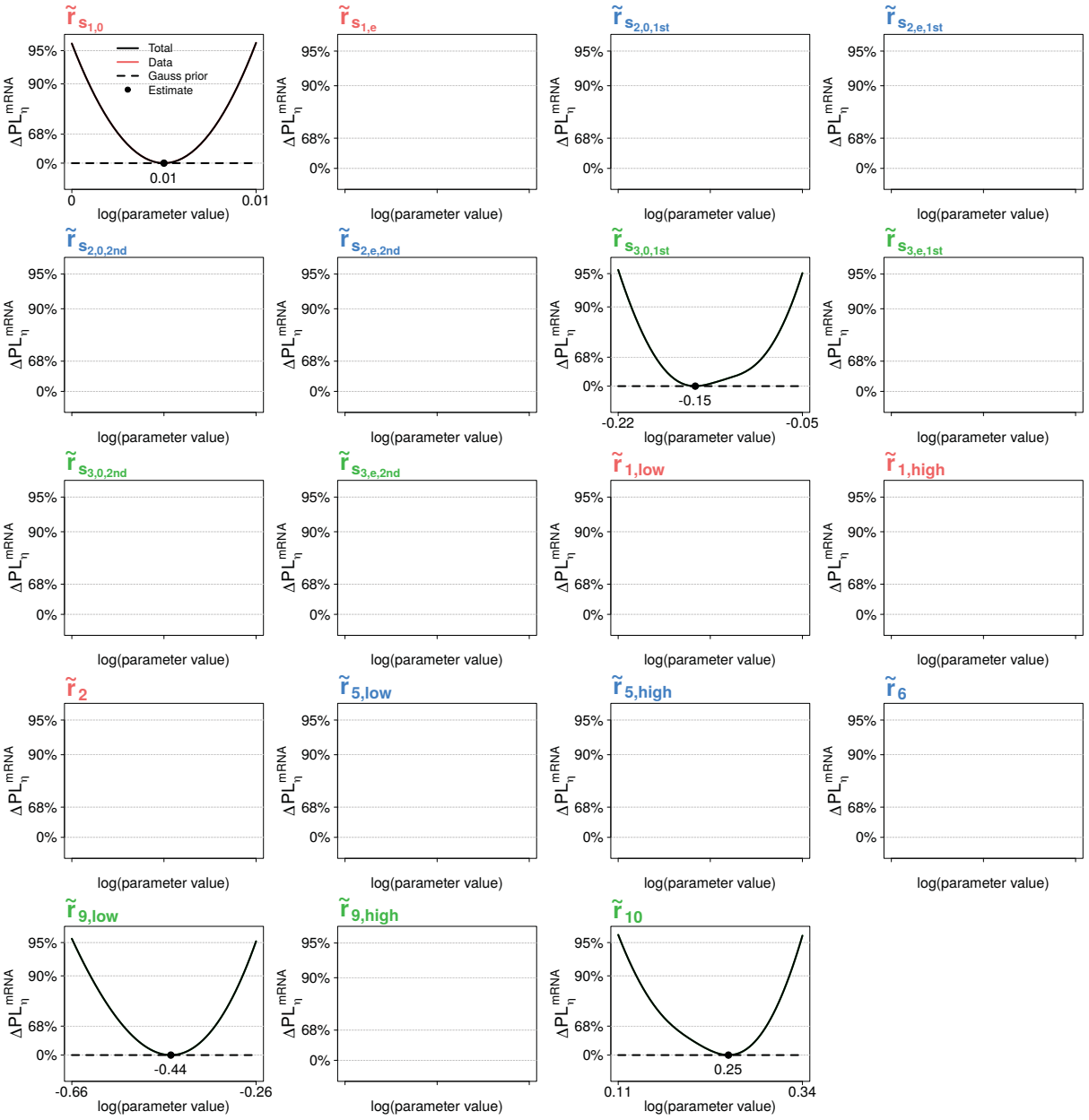

Figure E.34: Profile likelihoods of mRNA fold changes estimated for elutriated cells after L_1 regularization in the protein optimization step. In this figure, we represent profile likelihoods of mRNA fold changes estimated from Western blot data after applying L_1 regularization to mRNA fold changes in the protein optimization step as shown in Figure 7.1. mRNA fold changes compare mRNA parameters between unsynchronized und synchronized cells . Smallest, largest and estimated log-transformed parameter values are plotted on the x-axis. Likelihood ratios and different confidence levels are given on the y-axis. Colored lines show contributions of the data. Red, blue and green refer to Sic1, Cln2 and Clb5, respectively. The contribution of the Gaussian prior is given as dashed line and the total profile as black solid line. Parameter estimates are marked as black dots. An empty plot shows parameters which gone to zero during L_1 regularization.

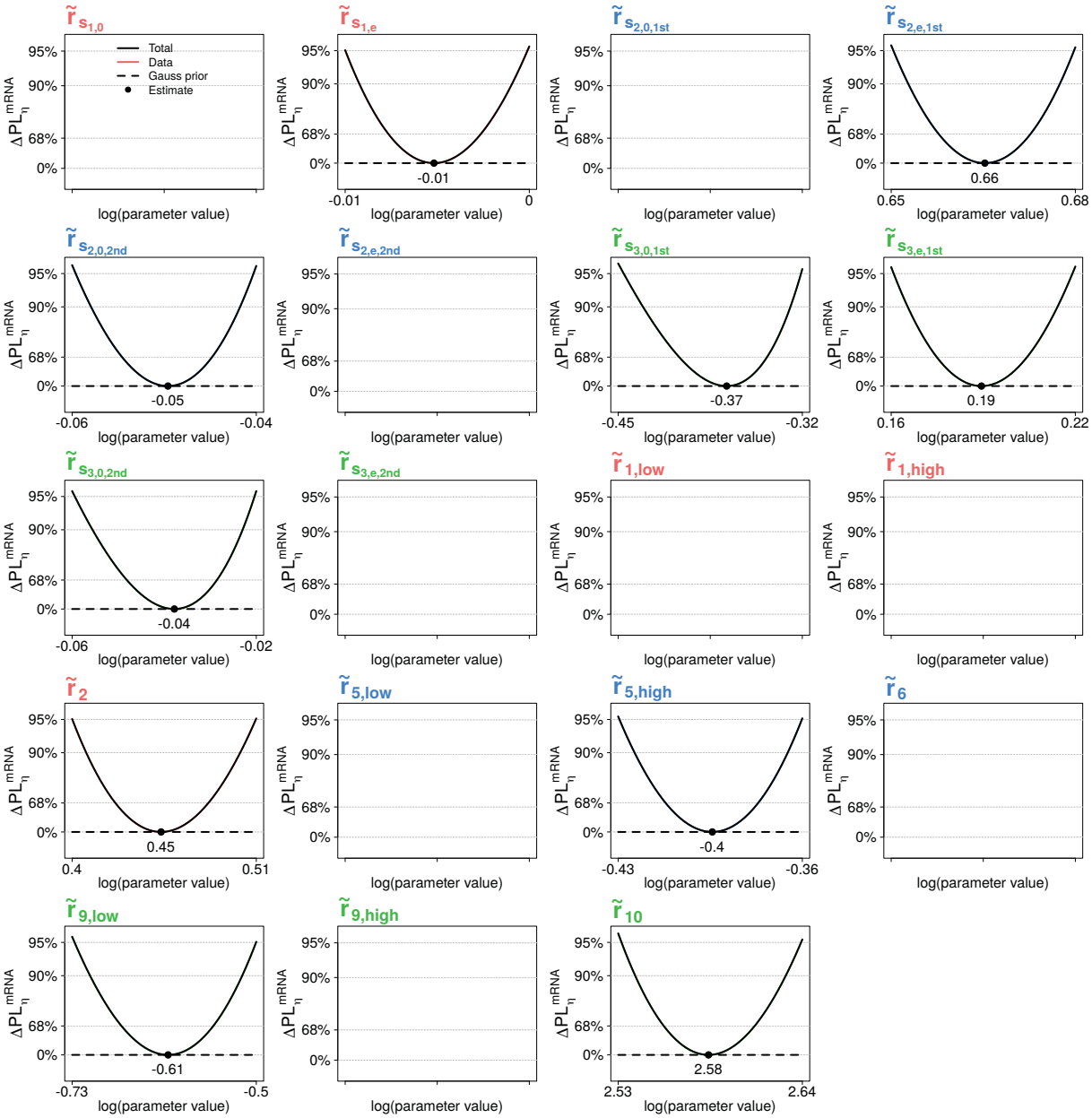

Figure E.35: Profile likelihoods of mRNA fold changes estimated after L_1 regularization for α-factor synchronized cells in the protein optimization step. In this figure, we represent profile likelihoods of mRNA fold changes estimated from Western blot data after applying L_1 regularization to mRNA fold changes in the protein optimization step as shown in Figure 7.1. mRNA fold changes compare mRNA parameters between unsynchronized und synchronized cells. Smallest, largest and estimated log-transformed parameter values are plotted on teh x-axis. Likelihood ratios and different confidence levels are given on the y-axis. Colored lines show contributions of the data. Red, blue and green refer to Sic1, Cln2 and Clb5, respectively. The contribution of the Gaussian prior is given as dashed line and the total profile as black solid line. Parameter estimates are marked as black dots. An empty plot shows parameters which gone to zero during L_1 regularization.

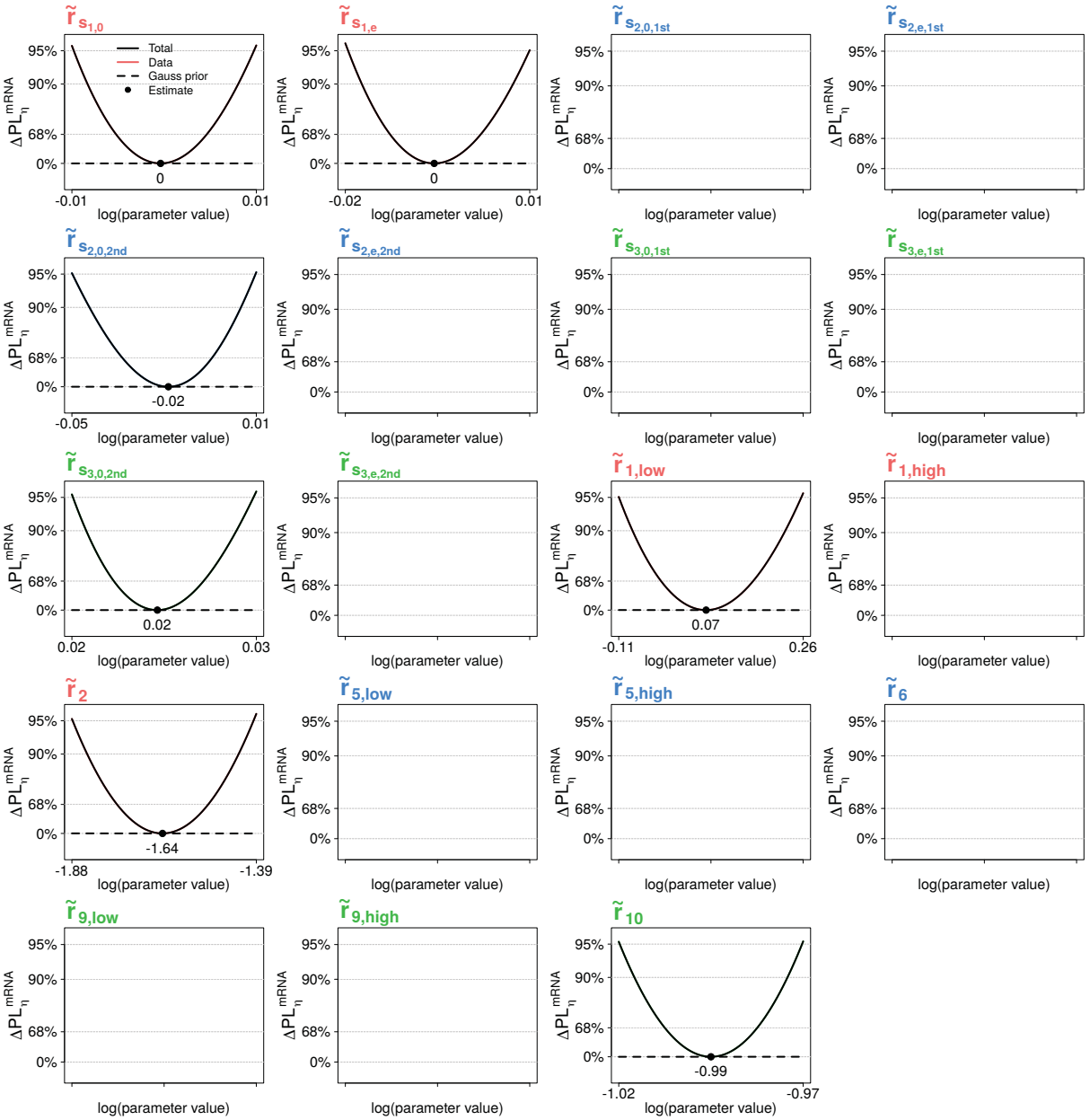

Figure E.36: Profile likelihoods of mRNA fold changes estimated after L_1 regularization for hydroxyurea synchronized cells in the protein optimization step. In this figure, we represent profile likelihoods of mRNA fold changes estimated from Western blot data after applying L_1 regularization to mRNA fold changes in the protein optimization step as shown in Figure 7.1. mRNA fold changes compare mRNA parameters between unsynchronized und synchronized cells. Smallest, largest and estimated log-transformed parameter values are plotted on the x-axis. Likelihood ratios and different confidence levels are given on the y-axis. Colored lines show contributions of the data. Red, blue and green refer to Sic1, Cln2 and Clb5, respectively. The contribution of the Gaussian prior is given as dashed line and the total profile as black solid line. Parameter estimates are marked as black dots. An empty plot shows parameters which gone to zero during L_1 regularization.

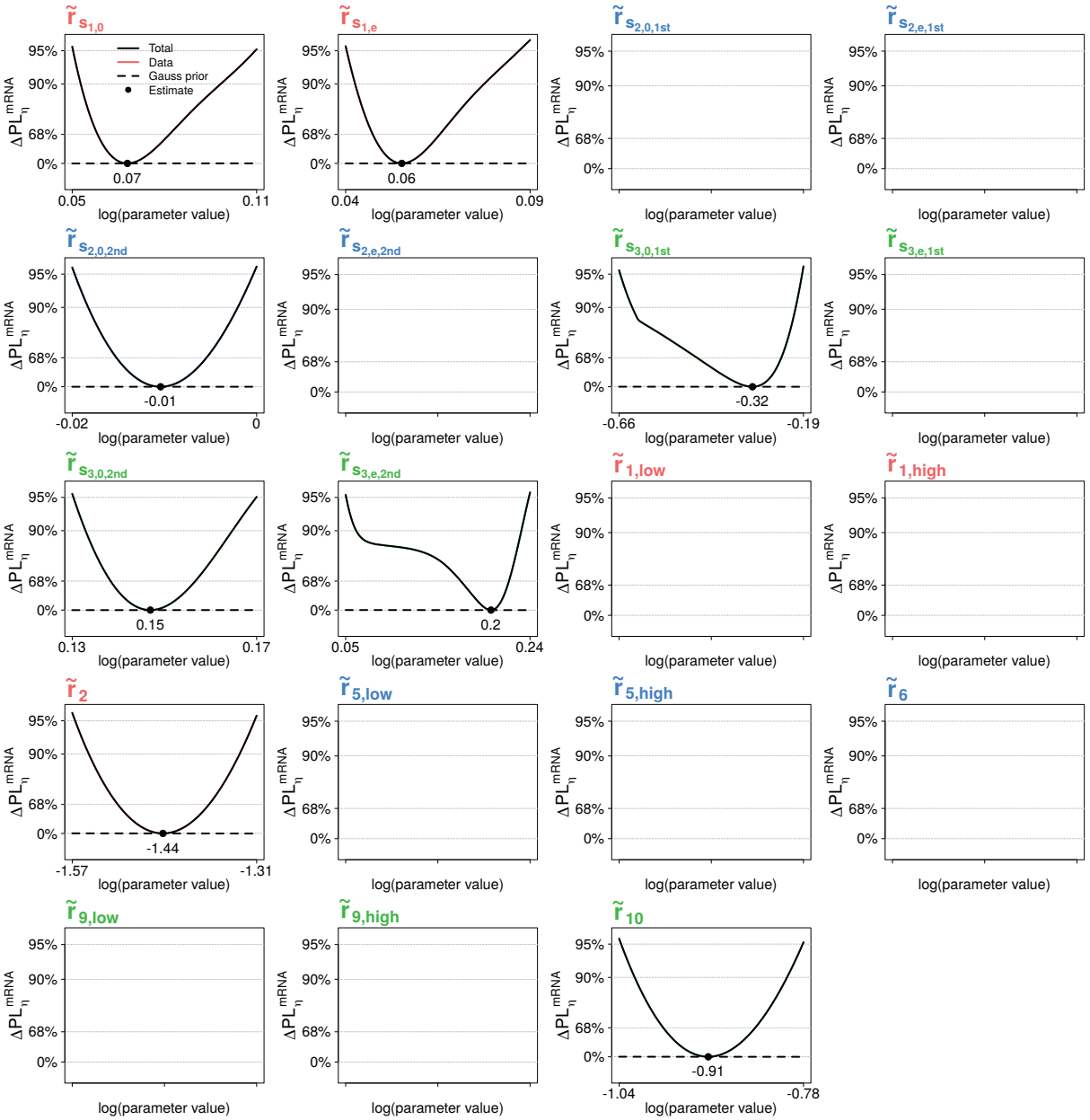

Figure E.37: Profile likelihoods of mRNA fold changes estimated after L_1 regularization for nocodazole synchronized cells in the protein optimization step. In this figure, we represent profile likelihoods of mRNA fold changes estimated from Western blot data after applying L_1 regularization to mRNA fold changes in the protein optimization step as shown in Figure 7.1. mRNA fold changes compare mRNA parameters between unsynchronized und synchronized cells. Smallest, largest and estimated log-transformed parameter values are plotted on the x-axis. Likelihood ratios and different confidence levels are given on the y-axis. Colored lines show contributions of the data. Red, blue and green refer to Sic1, Cln2 and Clb5, respectively. The contribution of the Gaussian prior is given as dashed line and the total profile as black solid line. Parameter estimates are marked as black dots. An empty plot shows parameters which gone to zero during L_1 regularization.

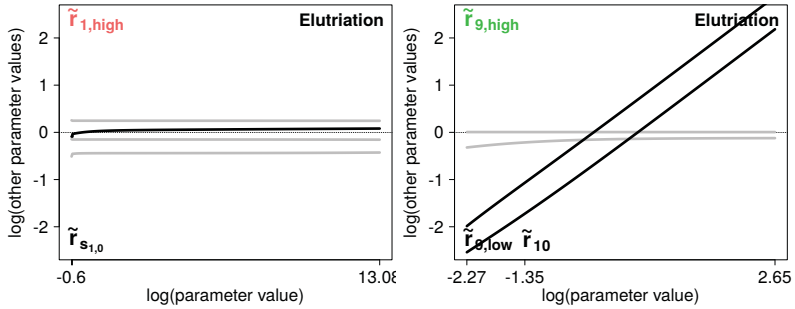

Figure E.38: Parameter values along profile likelihoods of L_1 removed mRNA fold changes for synchronization by elutriation. In this figure, we show parameter values along profile likelihoods of L_1 removed mRNA fold changes $\tilde{r}_{1,high}$ and $\tilde{r}_{9,high}$ which are subsequently re-introduced to test uniqueness of the solution. mRNA fold changes are log-transformed. Black lines indicate selected mRNA fold changes which are exchangeable with the non-selected mRNA fold change why the non-selected mRNA fold change is not uniquely determined to be zero. Exchangeable mRNA fold changes are noted in the plot. A detailed plot description is given in Figure 6.2.

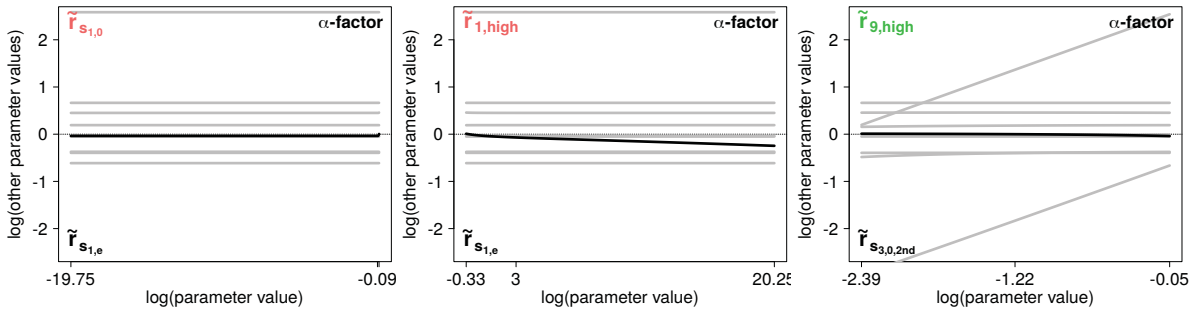

Figure E.39: Parameter values along profile likelihoods of L_1 removed mRNA fold changes for α-factor synchronization. In this figure, we show parameter values along profile likelihoods of L_1 removed mRNA fold changes $\tilde{r}_{S_{1,0}}$, $\tilde{r}_{1,high}$ and $\tilde{r}_{9,high}$ which are subsequently re-introduced to test uniqueness of the solution. mRNA fold changes are log-transformed. Black lines indicate selected mRNA fold changes which are exchangeable with the non-selected mRNA fold change why the non-selected mRNA fold change is not uniquely determined to be zero. Exchangeable mRNA fold changes are noted in the plot. A detailed plot description is given in Figure 6.2.

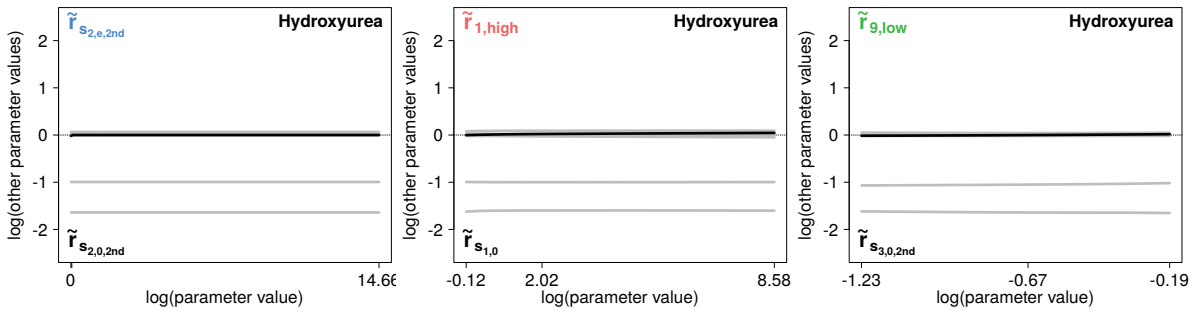

Figure E.40: Parameter values along profile likelihoods of L_1 removed mRNA fold changes for hydroxyurea synchronization. In this figure, we show parameter values along profile likelihoods of L_1 removed mRNA fold changes $\tilde{r}_{S_{2,e,2nd}}$, $\tilde{r}_{1,high}$ and $\tilde{r}_{9,low}$ which are subsequently re-introduced to test uniqueness of the solution. mRNA fold changes are log-transformed. Black lines indicate selected mRNA fold changes which are exchangeable with the non-selected mRNA fold change why the non-selected mRNA fold change is not uniquely determined to be zero. Exchangeable mRNA fold changes are noted in the plot. A detailed plot description is given in Figure 6.2.

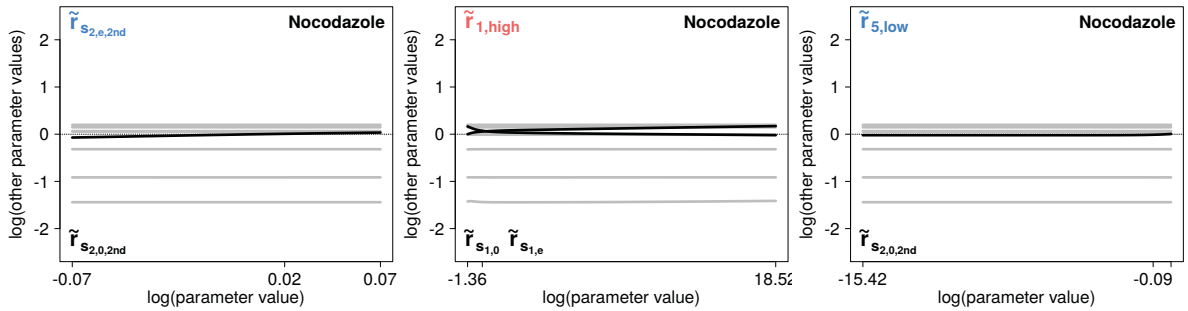

Figure E.41: Parameter values along profile likelihoods of L_1 removed mRNA fold changes for nocodazole synchronization. In this figure, we show parameter values along profile likelihoods of L_1 removed mRNA fold changes $\tilde{r}_{S_{2,e,2nd}}$, $\tilde{r}_{1,high}$ and $\tilde{r}_{5,low}$ which are subsequently re-introduced to test uniqueness of the solution. mRNA fold changes are log-transformed. Black lines indicate selected mRNA fold changes which are exchangeable with the non-selected mRNA fold change why the non-selected mRNA fold change is not uniquely determined to be zero. Exchangeable mRNA fold changes are noted in the plot. A detailed plot description is given in Figure 6.2.

F. Tables

$\boldsymbol{\nu}^T$	$X_1(t)$	$X_2(t)$	$X_3(t)$	$X_4(t)$	$X_5(t)$	$X_6(t)$	$X_7(t)$	$X_8(t)$	$X_9(t)$
R_1	+1	0	0	0	0	0	0	0	0
R_2	−1	0	0	0	0	0	0	0	0
R_3	−1	+1	0	0	0	0	0	0	0
R_4	0	−1	0	0	0	0	0	0	0
R_5	0	0	+1	0	0	0	0	0	0
R_6	0	0	−1	0	0	0	0	0	0
R_7	0	0	+1	−1	0	0	0	0	0
R_8	0	0	0	−1	0	0	0	0	0
R_9	0	0	0	0	+1	0	0	0	0
R_{10}	0	0	0	0	−1	0	0	0	0
R_{11}	0	0	0	0	−1	+1	0	0	0
R_{12}	0	0	0	0	0	0	−1	0	0
R_{13}	0	−1	0	0	0	−1	+1	0	0
R_{14}	0	0	0	−1	0	0	−1	+1	0
R_{15}	0	0	0	0	0	0	−1	+1	0
R_{16}	0	0	0	0	0	0	0	−1	+1
R_{17}	0	0	0	0	0	0	0	0	−1

Table F.1: Stoichiometric matrix of the chemical reaction system. This table shows the stoichiometric matrix of the chemical reaction system described in Table 4.2. Stoichiometric matrix $\boldsymbol{\nu}$ is a $N \times M$ matrix, here represented as transposed matrix $\boldsymbol{\nu}^T$. Each entry $\nu_{ji}^T = \nu_{ij}$ specifies the change in the number of molecules X_i of species S_i at time t caused by reaction R_j. These entries depend on the reaction type defined in Equation (4.2) and are given by the $N \times N$ identity maxtrix ϵ. State changes for inflow reactions (R_1, R_5 and R_9 corresponding to Equation (4.2a)) are calculated by $\nu_{ij} = \epsilon_{ii}$ and for outflow reactions (R_2, R_4, R_6, R_8, R_{10}, R_{12} and R_{16} corresponding to Equation (4.2b)) by $\nu_{ij} = -\epsilon_{ih}$. Conversion reactions ($R_3$, R_7, R_{11} and R_{15} corresponding to Equation (4.2c)) have state changes $\nu_{ij} = \epsilon_{ii} - \epsilon_{ih}$ and bimolecular reactions (R_{13}, R_{14} and R_{17} corresponding to Equation (4.2d)) $\nu_{ij} = \epsilon_{ii} - \epsilon_{ih} - \epsilon_{ig}$.

Par	No synchronization	Synchronization			
		Elutriation	α-factor	Hydroxyurea	Nocodazole
$t_{s_{1,0}}$	118.08	118.11	117.91	117.19	124.30
$t_{s_{1,e}}$	124.05	123.99	123.67	123.47	128.16
$t_{s_{2,0,1st}}$	20.48	19.87	21.90	20.45	19.60
$t_{s_{2,e,1st}}$	40.47	46.03	69.82	39.68	34.40
$t_{s_{2,0,2nd}}$	119.42	118.87	118.52	117.69	118.97
$t_{s_{2,e,2nd}}$	123.36	122.75	122.78	124.12	125.43
$t_{s_{3,0,1st}}$	40.16	35.77	26.32	40.11	26.59
$t_{s_{3,e,1st}}$	62.58	63.02	73.14	62.59	62.56
$t_{s_{3,0,2nd}}$	120.26	120.26	117.79	120.37	122.40
$t_{s_{3,e,2nd}}$	124.35	124.35	124.39	125.28	122.77
$k_{1,low}$	1.72	1.87	2.03	1.98	0.69
$k_{1,high}$	22.20	21.91	21.83	21.58	24.04
k_2	0.56	0.61	1.12	0.12	0.08
$k_{5,low}$	0.02	0.04	0.15	0.01	0.02
$k_{5,high}$	4.34	3.71	2.72	4.37	3.98
k_6	0.09	0.09	0.07	0.09	0.09
$k_{9,low}$	0.45	0.35	0.08	0.19	0.56
$k_{9,high}$	3.05	3.71	0.96	2.91	3.76
k_{10}	0.50	076	2.28	0.17	0.23
$scale$	-	0.51	1.72	1.02	1.57

Table F.2: mRNA parameter values. This table shows mRNA parameters estimated from smFISH data in the mRNA optimization step and mRNA parameters re-estimated from Western blot data in the protein optimization step. For reasons of consistency, the scaling factor introduced to combine mRNA and protein optimization is listed as well. Parameter values are rounded to the second decimal.

Par	Synchronization			
	Elutriation	α-factor	Hydroxyurea	Nocodazole
$C_{1,0}$	-	-	0.25	22.63
$C_{2,0}$	-	-	366.19	0,40
$C_{3,0}$	-	-	0.11	0.14
$C_{4,0}$	-	-	333.27	65.98
$C_{5,0}$	-	-	3048.15	158.96
$C_{6,0}$	-	-	0.37	330.01
$C_{7,0}$	-	-	0.38	0.45
$C_{8,0}$	-	-	139.17	29.75
$C_{9,0}$	-	-	0.35	0.30
k_3	1.10	20.43	2.56	1.24
k_4	0.005	0.08	0.12	0.03
k_7	0.49	3344.16	127.87	7.40
k_8	0.0002	74.65	0.17	0.006
k_{11}	1.79	150.27	5.26	68.90
k_{12}	0.09	0.0002	317.25	568.02
k_{13}	1.14	0.005	10.58	0.07
k_{14}	0.0008	0.001	0.49	0.45
k_{15}	0.05	0.27	0.19	0.06
k_{16}	0.006	26.59	0.29	4.59
k_{17}	0.001	9.76	0.42	0.39

Table F.3: Protein parameter before L_1 regularization. This table shows protein parameters estimated from Western blot data in the protein optimization step before applying L_1 regularization. Parameter values are rounded to the second decimal or the first decimal different from zero.

Par	Synchronization			
	Elutriation	α-factor	Hydroxyurea	Nocodazole
$C_{1,0}$	-	-	-	26.46
$C_{2,0}$	-	-	333.15	-
$C_{3,0}$	-	-	-	0.13
$C_{4,0}$	-	-	-	65.93
$C_{5,0}$	-	-	11201.53	-
$C_{6,0}$	-	-	-	330.04
$C_{7,0}$	-	-	-	-
$C_{8,0}$	-	-	137.29	28.64
$C_{9,0}$	-	-	-	-
k_3	1.12	15.46	3.03	1.24
k_4	0.004	0.06	0.15	0.02
k_7	0.44	39.82	187.05	7.35
k_8	0.0002	0.90	0.40	0.006
k_{11}	2.04	129.21	7.59	3.45
k_{12}	-	0.00007	285.75	0.45
k_{13}	-	-	-	-
k_{14}	-	-	-	-
k_{15}	0.04	0.19	0.22	0.09
k_{16}	0.01	-	0.36	-
k_{17}	-	-	-	-

Table F.4: Protein parameter after L_1 regularization. This table shows protein parameters estimated from Western blot data in the protein optimization step after applying L_1 regularization. Parameter values are rounded to the second decimal or the first decimal different from zero.

Group 1	Group 2	Group 3	Group 4	Group 5
$C_{1,0}$	$C_{3,0}$	$C_{5,0}$	$C_{3,0}$	$C_{1,0}$
$k_{1,low}$, $k_{1,high}$	$k_{5,low}$, $k_{5,high}$	$k_{9,low}$, $k_{9,high}$	$C_{4,0}$	$C_{2,0}$
k_3*	k_7*	$k_{11}*$	$k_{5,low}$, $k_{5,high}$	$C_{5,0}$
			$k_{14}*$	$C_{6,0}$
				$C_{7,0}$
				$C_{8,0}$
				$C_{9,0}$
				$k_{5,low}$, $k_{1,high}$
				$k_{5,low}$, $k_{9,high}$
				$k_{13}*$
				$k_{17}*$

Table F.5: Symmetry groups of the chemical reaction system. In this table, we show symmetry groups determined for the chemical reaction system (see Table 4.2) by using the R function `symmetryDetection()` of the R package **dMod** which is based on Lie-group symmetries. Symmetry detection is not possible for time dependent mRNA production rates defined by Fermi-Dirac distributions. The reason is that a Fermi-Dirac distribution includes an exponential function (see Appendix D.8) which is used in the R function to calculate scaling transformations. Thus, timing parameters are not included in the analysis and we can either have low or high production rates. Asterisks mark parameters which can be set to one to overcome scaling symmetries. Group 4 and Group 5 disappear due to observables $Y_2(t,\theta)$, $Y_4(t,\theta)$ and $Y_6(t,\theta)$ (see Table 5.1).

Bibliography 163

Acknowledgement 171

Declaration of authorship 173

Bibliography

[1] H. Karathia, E. Vilaprinyo, A. Sorribas, and R. Alves. Saccharomyces cerevisiae as a model organism: a comparative study. *PloS One*, 6(2):e16015, 2011.

[2] L. Pray. L. H. Hartwell's yeast: a model organism for studying somatic mutations and cancer. *Nature Education*, 1(1), 2008.

[3] A. Goffeau, B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, E. J. Louis, H. W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, and S. G. Oliver. Life with 6000 genes. *Science*, 274(5287):546–567, 1996.

[4] D. R. Zerbino, P. Achuthan, W. Akanni, M. R. Amode, D. Barrell, J. Bhai, K. Billis, C. Cummins, A. Gall, C. G. Girón, L. Gil, L. Gordon, L. Haggerty, E. Haskell, T. Hourlier, O. G. Izuogu, S. H. Janacek, T. Juettemann, J. K. To, M. R. Laird, I. Lavidas, Z. Liu, J. E. Loveland, T. Maurel, W. McLaren, B. Moore, J. Mudge, D. N. Murphy, V. Newman, M. Nuhn, D. Ogeh, C. K. Ong, A. Parker, M. Patricio, H. S. Riat, H. Schuilenburg, D. Sheppard, H. Sparrow, K. Taylor, A. Thormann, A. Vullo, B. Walts, A. Zadissa, A. Frankish, S. E. Hunt, M. Kostadima, N. Langridge, F. J. Martin, M. Muffato, E. Perry, M. Ruffier, D. M. Staines, S. J. Trevanion, B. L. Aken, F. Cunningham, A. Yates, and P. Flicek. Ensembl 2018. *Nucleic Acids Research*, 46(D1):D754–D761, 2018.

[5] D. Barik, D. A. Ball, J. Peccoud, and J. J. Tyson. A stochastic model of the yeast cell cycle reveals roles for feedback regulation in limiting cellular variability. *PLoS Computational Biology*, 12(12):e1005230, 2016.

[6] J. C. Mell and S. M. Burgess. Yeast as a model genetic organism. *Encyclopedia of Life Sciences*, 2002.

[7] L. H. Hartwell. Yeast and cancer. *Bioscience Reports*, 22(3-4):373–394, 2002.

[8] E. Klipp, W. Liebermeister, C. Wierling, A. Kowald, and R. Herwig. *Systems biology: a textbook*. John Wiley & Sons, 2016.

[9] K. Nasmyth. At the heart of the budding yeast cell cycle. *Trends in Genetics*, 12(10):405–412, 1996.

[10] L. H. Hartwell and M. W. Unger. Unequal devision in Saccharomyces cerevisiae and its implications for the control of cell division. *Journal of Cell Biology*, 75(11):433–435, 1977.

[11] J. J. Tyson and B. Novák. Temporal organization of the cell cycle. *Current Biology*, 18(17):759–768, 2008.

[12] J. J. Tyson and B. Novák. Regulation of the eukaryotic cell cycle: molecular antagonism, hysteresis, and irreversible transitions. *Journal of Theoretical Biology*, 210(2):249–263, 2001.

[13] S. J. Elledge. Cell cycle checkpoints preventing an identity crisis. *Science*, 274(5293):1664–1672, 1996.

[14] M. Malumbers and M. Barbacid. Cell cycle, CDKs and cancer: a changing paradigm. *Nature Reviews Cancer*, 9:153–166, 2015.

[15] K. Kaizu, S. Ghosh, Y. Matsuoka, H. Moriya, Y. Shimizu-Yoshida, and H. Kitano. A comprehensive molecular interaction map of the budding yeast cell cycle. *Molecular Systems Biology*, 6(415):1–11, 2010.

[16] C. Bertoli, J. M. Skotheim, and R. A. M. de Bruin. Control of cell cycle transcription during G1 and S phases. *Nature Reviews Molecular Cell Biology*, 14(8):518–528, 2013.

[17] E. Schwob, M. D. Mendenhall, and L. P. Markey. The B-type cyclin kinase inhibitor p40 controls the G1 to S transition in S. cerevisiae. *Cell*, 79(2):233–244, 1994.

[18] M. Kõivomägi, E. Valk, R. Venta, A. Iofik, M. Lepiku, D. O. Morgan, and M. Loog. Dynamics of Cdk1 substrate specificity during the cell cycle. *Molecular Cell*, 42(5):610–623, 2011.

[19] N. P. Gauthier, M. E. Larsen, R. Wernersson, U. de Lichtenberg, L. J. Jensen, S. Brunak, and T. S. Jensen. Cyclebase.org - a comprehensive multi-organism online database of cell-cycle experiments. *Nucleic Acids Research*, 36:854–859, 2008.

[20] N. P. Gauthier, L. J. Jensen, R. Wernersson, S. Brunak, and T. S. Jensen. Cyclebase.org: version 2.0, an updated comprehensive, multi-species repository of cell cycle experiments and derived analysis results. *Nucleic Acids Research*, 38:699–702, 2009.

[21] A. Santos, R. Wernersson, and L. J. Jensen. Cyclebase 3.0: a multi-organism database on cell-cycle regulation and phenotypes. *Nucleic Acids Research*, 43(D1):D1140–D1144, 2014.

[22] M. Kõivomägi, E. Valk, R. Venta, A. Iofik, M. Lepiku, E. R. M. Balog, S. M. Rubin, D. O. Morgan, and M. Loog. Cascades of multisite phosphorylation control Sic1 destruction at the onset of S phase. *Nature*, 480(7375):128–131, 2011.

[23] A. Campbell. Synchronization of cell devision. *Bacteriology Reviews*, 21(4):263–272, 1957.

[24] S. M. Jazwinski. Aging and senescence of the budding yeast Saccharomyces cerevisiae. *Molecular Microbiology*, 4(3):337–343, 1990.

[25] K. K. Steffen, B. K. Kennedy, and M. Kaeberlein. Measuring replicative life span in the budding yeast. *Journal of Visualized Experiments*, 28:e1209, 2009.

[26] R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, David Landsman, David J. Lockhart, and Ronald W. Davis. A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell*, 2(1):65–73, 1998.

[27] P. T. Spellman, G. Sherlock, M. Q. Zhang, R. Vishwanath, K. Anders, M. B. Eisen, P. O. Brown, and B. Futcher. Comprehensive identification of cell cycle–regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9(12):3273–3297, 1998.

[28] B. Futcher. Cell cycle synchronization. *Methods in Cell Science*, 21(2-3):79–86, 1999.

[29] K. Kuritz, D. Stöhr, N. Pollak, and F. Allgöwer. On the relationship between cell cycle analysis with ergodic principles and age-structured cell population models. *Journal of Theoretical Biology*, 414:91–102, 2017.

[30] A. Balanov, N. Janson, D. Postnov, and O. Sosnovtseva. *Synchronization: from simple to complex*. Springer, 2008.

[31] F. R Cross. Starting the cell cycle : what's the point? *Current Opinion in Cell Biology*, 7(6):790–797, 1995.

[32] A. Day, C. Schneider, and B. L. Schneider. Yeast cell synchronization. In *Cell Cycle Checkpoint Control Protocols*, pages 55–76. Humana Press, 2004.

[33] S. J. Elledge, Z. Zhou, J. B. Allen, and T. A. Navas. DNA damage and cell cycle regulation of ribonucleotide reductase. *BioEssays*, 15(5):333–339, 1993.

[34] C. W. Jacobs, A. E. Adams, P. J. Szaniszlo, and J. R. Pringle. Functions of microtubules in the Saccharomyces cerevisiae cell cycle. *Journal of Cell Biology*, 107(4):1409–1426, 1988.

[35] A. S. Zhuk, E. I. Stepchenkova, Y. I. Pavlov, and S. G. Inge-Vechtomov. Evaluation of effectiveness of synchronization methods of cell division in yeast Saccharomyces cerevisiae. *Tsitologiya*, 58(12):936–946, 2016.

[36] J. Y. Hur, M. C. Park, K. Y. Suh, and S. H. Park. Synchronization of cell cycle of saccharomyces cerevisiae by using a cell chip platform. *Molecules and Cells*, 32(5):483–488, 2011.

[37] G. M. Walker. Synchronization of yeast cell populations. *Methods in Cell Science*, 21(2-3):87–93, 1999.

[38] S. J. Altschuler and L. F. Wu. Cellular heterogeneity: do differences make a difference? *Cell*, 141(4):559–563, 2010.

[39] N. Samusik, N. Aghaeepour, and S. Bendall. Single-cell analysis and modelling of cell population heterogeneity. In *Pacific Symposium on Biocomputing 2017*, pages 557–564, 2017.

[40] B. Munsky. Using gene expression noise to understand gene regulation. *Science*, 336(6078):183–188, 2012.

[41] M. B. Elowitz, A.J. Levine, E. D. Siggia, and P. S. Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186., 2002.

[42] J. M. Raser. Noise in gene expression: origins, consequences and control. *Science*, 309(5743):2010–2013, 2010.

[43] J. M. Raser and E. O'Shea. Control of stochasticity in eukaryotic gene expression. *Science*, 1811(2004):1811–1815, 2011.

[44] A. Llamosi, A. M. Gonzalez-Vargas, C. Versari, E. Cinquemani, G. Ferrari-Trecate, P. Hersen, and G. Batt. What population reveals about individual cell identity: single-cell parameter estimation of models of gene expression in yeast. *PLoS Computational Biology*, 12(2):e1004706, 2016.

[45] A. Raj and A. van Oudenaarden. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, 135(2):216–226, 2008.

[46] John R.S. Newman, Sina Ghaemmaghami, Jan Ihmels, David K. Breslow, Matthew Noble, Joseph L. DeRisi, and Jonathan S. Weissman. Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature*, 441(7095):840–846, 2006.

[47] N. Navin, J. Kendall, J. Troge, P. Andrews, L. Rodgers, J. McIndoo, K. Cook, A. Stepansky, D. Levy, D. Esposito, L. Muthuswamy, A. Krasnitz, W. R. McCombie, J. Hicks, and M. Wigler. Tumour evolution inferred by single-cell sequencing. *Nature*, 472(7341):90–95, 2011.

[48] D. A. Jaitin, E. Kenigsberg, H. Keren-Shaul, N. Elefant, F. Paul, I. Zaretsky, A. Mildner, N. Cohen, S. Jung, A. Tanay, and I. Amit. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science*, 343(6172):776–779, 2014.

[49] L. Teufel, K. Tummler, M. Flöttmann, A. Herrmann, N. Barkai, and E. Klipp. Roles of G1 cyclins in the temporal organization of yeast cell cycle - a transcriptome-wide analysis. *bioRxiv*, 2018.

[50] M. À. Adrover, Z. Zi, A. Duch, J. Schaber, A. González-Novo, J. Jimenez, Ma. Nadal-Ribelles, J. Clotet, E. Klipp, and F. Posas. Time-dependent quantitative multicomponent control of the G1-S network by the stress-activated protein kinase Hog1 upon osmostress. *Science Signaling*, 4(192):ra63, 2011.

[51] A. J. Hughes, D. P. Spelke, Z. Xu, C.-C. Kang, D. V. Schaffer, and A. E. Herr. Single-cell western blotting. *Nature Methods*, 11(7):455–464, 2015.

[52] A. Amoussouvi, L. Teufel, M. Reis, M. Seeger, J. K. Schlichting, G. Schreiber, A. Herrmann, and E. Klipp. Transcriptional timing and noise of yeast cell cycle regulators – a single cell and single molecule approach. *npj Systems Biology and Applications*, 4(17):1–10, 2018.

[53] A. P. Frei, F. A. Bava, E. R. Zunder, E. W. Y. Hsieh, S. Yu Chen, G.P. Nolan, and P. F. Gherardini. Highly multiplexed simultaneous detection of RNAs and proteins in single cells. *Nature Methods*, 13(3):269–275, 2016.

[54] A. Lyubimova, S. Itzkovitz, J. P. Junker, Z. P. Fan, X. Wu, and A. Van Oudenaarden. Single-molecule mRNA detection and counting in mammalian tissue. *Nature Protocols*, 8(9):1743–1758, 2013.

[55] J. H. Lee, E. R. Daugharthy, J. Scheiman, R. Kalhor, T. C. Ferrante, R. Terry, B. M. Turczyk, J. L. Yang, H. S. Lee, J. Aach, K. Zhang, and G. M. Church. Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nature Protocols*, 10(3):442–458, 2015.

[56] U. Abraham, J. K. Schlichting, A. Kramer, and H. Herzel. Quantitative analysis of circadian single cell oscillations in response to temperature. *PloS One*, 13(1):e0190004, 2018.

[57] H. Kitano. Systems biology: a brief overview. *Science*, 295(5560):1662–1664, 2002.

[58] A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–1929, 2009.

[59] G. Cedersund. Conclusions via unique predictions obtained despite unidentifiability - new definitions and a general method. *FEBS Journal*, 279(18):3513–3527, 2012.

[60] J. Sun, J. M. Garibaldi, and C. Hodgman. Parameter estimation using metaheuristics in systems biology: a comprehensive review. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(1):185–202, 2012.

[61] R. Barlow. *Statistics: a guide to the use of statistical methods in the physical sciences*. John Wiley & Sons, 1989.

[62] A. Raue, B. Steiert, M. Schelker, C. Kreutz, T. Maiwald, H. Hass, J. Vanlier, C. Tönsing, L. Adlung, R. Engesser, W. Mader, T. Heinemann, J. Hasenauer, M. Schilling, T. Höfer, E. Klipp, F. Theis, U. Klingmüller, B. Schöberl, and J. Timmer. Data2Dynamics: a modeling environment tailored to parameter estimation in dynamical systems. *Bioinformatics*, 31(21):3558–3560, 2015.

[63] D. Kaschek, W. Mader, M. Rosenblatt, and J. Timmer. Dynamic modeling , parameter estimation and uncertainty analysis in R. *bioRxiv*, 2016.

[64] E. Balsa-Canto, D. Henriques, A. Gábor, and J. R. Banga. AMIGO2, a toolbox for dynamic modeling, optimization and control in systems biology. *Bioinformatics*, 32(21):3357–3359, 2016.

[65] J. A. Egea, D. Henriques, T. Cokelaer, A. F. Villaverde, A. MacNamara, D.-P. Danciu, J. R. Banga, and J. Saez-Rodriguez. MEIGO: an open-source software suite based on metaheuristics for global optimization in systems biology and bioinformatics. *BMC Bioinformatics*, 15(1):136, 2014.

[66] P. Stapor, D. Weindl, B. Ballnus, S. Hug, C. Loos, A. Fiedler, S. Krause, S. Hroß, F. Fröhlich, and J. Hasenauer. PESTO: Parameter EStimation TOolbox. *Bioinformatics*, 34(4):705–707, 2018.

[67] S. Hoops, R. Gauges, C. Lee, J. Pahle, N. Simus, M. Singhal, L. Xu, P. Mendes, and U. Kummer. COPASI - A COmplex PAthway SImulator. *Bioinformatics*, 22(24):3067–3074, 2006.

[68] Z. Zi and E. Klipp. SBML-PET: a Systems Biology Markup Language-based parameter estimation tool. *Bioinformatics*, 22(21):2704–2705, 2006.

[69] A. I. Goranov, M. Cook, M. Ricicova, G. Ben-Ari, C. Gonzalez, C. Hansen, M. Tyers, and A. Amon. The rate of cell growth is governed by cell cycle stage. *Genes and Development*, 23(12):1408–1422, 2009.

[70] G. Neuert, B. Munsky, R. Z. Tan, L. Teytelman, M. Khammash, and A. van Oudenaarden. Systematic identification of signal-activated stochastic gene regulation. *Science*, 339(6119):584–587, 2013.

[71] J. Timmer. Parameter estimation in nonlinear stochastic differential equations. *Chaos, Solitons & Fractals*, 11(15):2571–2578, 2000.

[72] T. Trcek, D. R. Larson, A. Moldón, C. C. Query, and R. H. Singer. Single-molecule mRNA decay measurements reveal promoter-regulated mRNA stability in yeast. *Cell*, 147(7):1484–1497, 2011.

[73] T. Trcek, J. A. Chao, D. R. Larson, H. Y. Park, D. Zenklusen, S. M. Shenoy, and R. H. Singer. Single-mRNA counting using fluorescent in situ hybridization in budding yeast. *Nature Protocols*, 7(2):408–419, 2012.

[74] D. Zenklusen, D. R. Larson, and R. H. Singer. Single-RNA counting reveals alternative modes of gene expression in yeast. *Nature Structural and Molecular Biology*, 15(12):1263–1271, 2008.

[75] M. Costanzo, J. L. Nishikawa, X. Tang, J. S. Millman, O. Schub, K. Breitkreuz, D. Dewar, I. Rupes, B. Andrews, and M. Tyers. CDK activity antagonizes Whi5, an inhibitor of G1/S transcription in yeast. *Cell*, 117(7):899–913, 2004.

[76] C. Kreutz and J. Timmer. Systems biology: experimental design. *FEBS Journal*, 276(4):923–942, 2009.

[77] A. Raue, M. Schilling, J. Bachmann, A. Matteson, M. Schelke, D. Kaschek, S. Hug, C. Kreutz, B. D. Harms, F. J. Theis, U. Klingmüller, and J. Timmer. Lessons learned from quantitative dynamical modeling in systems biology. *PLoS One*, 8(9):e74335, 2013.

[78] A. Oppelt, D. Kaschek, S. Huppelschoten, R. Sison-Young, F. Zhang, M. Buck-Wiese, F. Herrmann, S. Malkusch, C. L. Krüger, M. Meub, B. Merkt, L. Zimmermann, A. Schofield, R. P. Jones, H. Malik, M. Schilling, M. Heilemann, B. van de Water, C. E. Goldring, B. K. Park, J. Timmer, and U. Klingmüller. Model-based identification of TNFα-induced IKKβ-mediated and IκBα-mediated regulation of NFκB signal transduction as a tool to quantify the impact of drug-induced liver injury compounds. *npj Systems Biology and Applications*, 4(23):1–16, 2018.

[79] C. Kreutz, M. M. Bartolome Rodriguez, T. Maiwald, M. Seidl, H. E. Blum, L. Mohr, and J. Timmer. An error model for protein quantification. *Bioinformatics*, 23(20):2747–2753, 2007.

[80] S. Hug, A. Raue, J. Hasenauer, J. Bachmannd, U. Klingmüller, J. Timmer, and F. J. Theis. High-dimensional Bayesian parameter estimation: case study for a model of JAK2/STAT5 signaling. *Mathematical Biosciences*, 246(2):293–304, 2013.

[81] M. Wang, M. Weiss, M. Simonovic, G. Haertinger, S. P. Schrimpf, M. O. Hengartner, and C. von Mering. PaxDb, a database of protein abundance averages across all three domains of life. *Molecular & Cellular Proteomics*, 11(8):492–500, 2012.

[82] S. Ghaemmaghami, W.-K. Huh, K. Bower, R. W. Howson, A. Belle, N. Dephoure, E. K. O'Shea, and J. S. Weissman. Global analysis of protein expression in yeast. *Nature*, 425(6959):737–741, 2003.

[83] C. Gardiner. *Stochastic methods*. Springer, 2009.

[84] D. T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403–434, 1976.

[85] D. T. Gillespie. A rigorous derivation of the chemical master equation. *Physica A: Statistical Mechanics and its Applications*, 188(1-3):404–425, 1992.

[86] D. T. Gillespie. Exact stochastic simulation of couple chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

[87] D. T. Gillespie. Stochastic simulation of chemical kinetics. *Annual Review of Physical Chemistry*, 58:35–55, 2007.

[88] D. T. Gillespie. Chemical Langevin equation. *Journal of Chemical Physics*, 113(1):297–306, 2000.

[89] D. J. Higham. Modeling and simulating chemical reactions. *SIAM Review*, 50(2):347–368, 2008.

[90] D. A. McQuarrie. Stochastic approach to chemical kinetics. *Applied Probability Trust*, 4(3):413–478, 1967.

[91] S. M. Ross. *Introduction to probability models*. Elsevier, 1950.

[92] N. G. Van Kampen. *Stochastic processes in physics and chemistry*. Elsevier, 2007.

[93] D. J. Wilkinson. *Stochastic modelling for systems biology*. CRC press, 2011.

[94] M. Ahmadian, S. Wang, and J. J. Tyson. Hybrid ODE / SSA model of the budding yeast cell cycle control mechanism with mutant case study. In *ACM-BCB 2017*, pages 464–473, 2017.

[95] T. Jahnke and W. Huisinga. Solving the chemical master equation for monomolecular reaction systems analytically. *Journal of Mathematical Biology*, 54(1):1–26, 2007.

[96] A. Azzalini. *Statistical inference based on the likelihood*. Chapman & Hall, 1996.

[97] F. Fröhlich, B. Kaltenbacher, F. J. Theis, and J. Hasenauer. Scalable parameter estimation for genome-scale biochemical reaction networks. *PLoS Computational Biology*, 13(1):e1005331., 2017.

[98] Y. Pawitan. *In all likelihood: statistical modelling and inference using likelihood*. Oxford University Press, 2001.

[99] J. R. Banga. Optimization in computational systems biology. *BMC Systems Biology*, 2(1):1–7, 2008.

[100] J. Nocedal and S. J. Wright. *Numerical optimization*. Springer, 2006.

[101] C G Broyden. The convergence of a class of double-rank minimization algorithms: 1. general considerations. *IMA Journal of Applied Mathematics*, 6(1):76–90, 1970.

[102] C . G . Broyden. The convergence of a class of double-rank minimization algorithms: 2. the new algorithm. *IMA Journal of Applied Mathematics*, 6(3):222–231, 1970.

[103] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

[104] A. Raue, C. Kreutz, T. Maiwald, U. Klingmüller, and J. Timmer. Addressing parameter identifiability by model-based experimentation. *IET Systems Biology*, 5(2):120–130, 2011.

[105] T. Heinemann and A. Raue. Model calibration and uncertainty analysis in signaling networks. *Current Opinion in Biotechnology*, 39:143–149, 2016.

[106] A. Raue, C. Kreutz, F. J. Theis, and J. Timmer. Joining forces of Bayesian and frequentist methodology: A study for inference in the presence of non-identifiability. *Philosophical Transactions of the Royal Society A*, 371(1984), 2013.

[107] J. Vanlier, C. A. Tiemann, P. A. J. Hilbers, and N. A. W. van Riel. An integrated strategy for prediction uncertainty analysis. *Bioinformatics*, 28(8):1130–1135, 2012.

[108] Y. Yuan. A review of trust region algorithms for pptimization. *Iciam*, 99:271–282, 2000.

[109] M. Peifer and J. Timmer. Parameter estimation in ordinary differential equations for biochemical processes using the method of multiple shooting. *IET Systems Biology*, 1(2):78–88, 2007.

[110] A. Raue, V. Becker, U. Klingmüller, and J. Timmer. Identifiability and observability analysis for experimental design in nonlinear dynamical models. *Chaos*, 20(4), 2010.

[111] C. Kreutz, A. Raue, D. Kaschek, and J. Timmer. Profile likelihood in systems biology. *FEBS Journal*, 280(11):2564–2571, 2013.

[112] S. Hengl, C. Kreutz, J. Timmer, and T. Maiwald. Data-based identifiability analysis of non-linear dynamical models. *Bioinformatics*, 23(19):2612–2618, 2007.

[113] B. Steiert, A. Raue, J. Timmer, and C. Kreutz. Experimental design for parameter estimation of gene regulatory networks. *PLoS One*, 7(7):e40052, 2012.

[114] O. Demir-Kavuk, M. Kamada, T. Akutsu, and E.-W. Knapp. Prediction using step-wise L1, L2 regularization and feature selection for small data sets with large number of features. *BMC Bioinformatics*, 12(1):412, 2011.

[115] B. Steiert, J. Timmer, and C. Kreutz. L1 regularization facilitates detection of cell type-specific parameters in dynamical systems. *Bioinformatics*, 32(17):i718–i726, 2016.

[116] T. Maiwald, H. Hass, B. Steiert, J. Vanlier, R. Engesser, A. Raue, F. Kipkeew, H. H. Bock, D. Kaschek, C. Kreutz, and J. Timmer. Driving the model to its limit: profile likelihood based model reduction. *PLoS One*, 11(9):e0162366, 2016.

[117] G. Cumming and S. Finch. Inference by eye confidence intervals and how to read pictures of data. *American Psychologist*, 60(2):170–180, 2005.

[118] D. Ball, N. R. Adames, N. Reischmann, D. Barik, C. T. Franck, J. J. Tyson, and J. Peccoud. Measurement and modeling of transcriptional noise in the cell cycle regulatory network. *Cell Cycle*, 12(19):3203–3218, 2013.

[119] J. Hasenauer, C. Hasenauer, T. Hucho, and F. J. Theis. ODE constrained mixture modelling: a method for unraveling subpopulation structures and dynamics. *PLoS Computational Biology*, 10(7), 2014.

[120] C. Loos, A. Fiedler, and J. Hasenauer. Parameter estimation for reaction rate equation constrained mixture models. In *CMSB 2016*, pages 186–200, 2016.

[121] L. Hu, X. Yin, J. Sun, A. Zetterberg, W. Miao, and T. Cheng. A molecular pathology method for sequential fluorescence in situ hybridization for multi-gene analysis at the single-cell level. *Oncotarget*, 8(31):50534–50541, 2016.

[122] J. Aach and G. M. Church. Aligning gene expression time series with time warping algorithms. *Bioinformatics*, 17(6):495–508, 2001.

[123] A. W. C. Fu, E. Keogh, L. Y. H. Lau, C. A. Ratanamahatana, and R. C. W. Wong. Scaling and time warping in time series querying. *VLDB Journal*, 17(4):899–921, 2008.

[124] S. Di Talia, J. M. Skotheim, J. M. Bean, E. D. Siggia, and F. R. Cross. The effects of molecular noise and size control on variability in the budding yeast cell cycle. *Nature*, 448(7156):947–951, 2007.

[125] B. Novák, A. Tóth, A. Csikász-Nagy, B. Gyorffy, J. J. Tyson, and K. Nasmyth. Finishing the cell cycle. *Journal of Theoretical Biology*, 199(2):223–233, 1999.

[126] K. C. Chen, A. Csikasz-Nagy, B. Gyorffy, J. Val, B. Novák, and J. J. Tyson. Kinetic analysis of a molecular model of the budding yeast cell cycle. *Molecular Biology of the Cell*, 11(1):369–391, 2000.

[127] K. C. Chen, L. Calzone, A. Csikasz-Nagy, F. R. Cross, B. Novák, and J. J. Tyson. Integrative analysis of cell cycle control in budding yeast. *Molecular Biology of the Cell*, 15(8):3751–3737, 2004.

[128] P. Kraikivski, K. C. Chen, T. Laomettachit, T. M. Murali, and J. J. Tyson. From START to FINISH: computational analysis of cell cycle control in budding yeast. *npj Systems Biology and Applications*, 1:15016, 2015.

[129] K. L. Lai and J. L. Crassidis. Extensions of the first and second complex-step derivative approximations. *Journal of Computational and Applied Mathematics*, 219(1):276–293, 2008.

[130] J. R. R. A. Martins, P. Sturdza, and J. J. Alonso. The complex-step derivative approximation. *ACM Transactions on Mathematical Software*, 29(3):245–262, 2003.

[131] F. Fröhlich, F. J. Theis, J. O. Rädler, and J. Hasenauer. Parameter estimation for dynamical systems with discrete events and logical operations. *Bioinformatics*, 33(7):1049–1056, 2017.

[132] G. Cumming. Inference by eye: reading the overlap of independent confidence intervals. *Statistics in Medicine*, 28:221–239, 2009.

[133] R. W. Smith. Visual hypothesis testing with confidence intervals. In *Proceedings of the Twenty-Second Annual SAS® Users Group International Conference*, pages 1–6, 1997.

[134] A. van Oudenaarden. Stochastic chemical kinetics. 2009.

[135] M. D. McKay, R. J. Beckman, and W. J. Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21(2):239–245, 1979.

Acknowledgement

In the last three years, many people supported me personally or supported my studies. My special thanks goes to all of them.

At first, I want to thank my supervisor **Edda Klipp**. I was always free to develop and to apply my own ideas and she always had an open door. Because of her, I was able to attend a number of meetings, workshops and conferences. Whenever I had an extra challenge, it was not a challenge for Edda and she found a solution immediately. Thanks to Edda and my graduate program "Computational System Biology", it was most of the time possible to keep the work-life balance.

Without any experimental data it is impossible to mathematically model biological processes. Therefore, I thank the experimentalists that measured the data represented in this study, especially **Gabriele Schreiber**. She answered all my naive questions, showed me experimental set ups and spent a lot of time to generate adequate Western blot data.

Many thanks belong to my proofreaders **Hugo Schlichting** and **Roman Rainer**. For years, I had many in-depth physical conversations with my father in his ivory tower. Roman was my officemate and always cooled me down if it was required, e.g. driving a car about 900 km over night from Freiburg to Berlin after the flight was canceled. Lately, he rescued my life, I had synchronization problems with my cloud and I thought I had overwritten my thesis. In general, every of my R and LaTex problems were solved by him. **Ivo Maintz** had the pleasure to solve my synchronization problem as well. I also want to thank him for his support, especially early in the morning. Everybody who knows my father knows that synchronization problems are common in my family.

Further, I have to thank Mathias Legrand for sharing his LaTex template[10] which was modified by myself.

I also want to thank **Hanspeter Herzel** and **Grigory Bordyugov**. They were not directly involved in this thesis but gave me many supports during my bachelor and master studies and thus enabled me to start my PhD.

My final thanks belongs to my mother **Sylvia Schlichting**, my husband **Sebastian Mattukat** and my whole family. Since my daughter was born, my mother took care of her whenever it was needed. I never had to take a day off if the child care was closed or my daughter was sick. When I was on a business trip, my mother and my husband managed her every day demands, from a full diaper up to leisure activities.

[10] downloaded from `http://www.LaTeXTemplates.com`

Declaration of authorship

I hereby declare that I completed the doctoral thesis independently based on the stated resources and aids. I have not applied for a doctoral degree elsewhere and do not have a corresponding doctoral degree. I have not submitted the doctoral thesis, or parts of it, to another academic institution and the thesis has not been accepted or rejected. I declare that I have acknowledged the Doctoral Degree Regulations which underlie the procedure of the Faculty of Life Sciences of Humboldt-Universität zu Berlin, as amended on 5th March 2015. Furthermore, I declare that no collaboration with commercial doctoral degree supervisors took place, and that the principles of Humboldt-Universität zu Berlin for ensuring good academic practice were abided by.

Berlin, December 4, 2018

Julia Katharina Schlichting